The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Cornell University

February 5, 2018
Examples of exploration and exploitation in real life

1. Going to your favorite restaurant/bar vs. trying a new one.
Examples of exploration and exploitation in real life

1. Going to your favorite restaurant/bar vs. trying a new one.
2. Listening to music from a band you love vs. discovering new ones.
Examples of exploration and exploitation in real life

1. Going to your favorite restaurant/bar vs. trying a new one.
2. Listening to music from a band you love vs. discovering new ones.
3. Preparing a meal that you have made successfully in the past and you enjoyed vs. cooking up something new.
Examples of exploration and exploitation in real life

1. Going to your favorite restaurant/bar vs. trying a new one.
2. Listening to music from a band you love vs. discovering new ones.
3. Preparing a meal that you have made successfully in the past and you enjoyed vs. cooking up something new.
4. Reading a newspaper article from a journalist you like vs. reading something from a newcomer in the field.
Examples of exploration and exploitation in real life

1. Going to your favorite restaurant/bar vs. trying a new one.
2. Listening to music from a band you love vs. discovering new ones.
3. Preparing a meal that you have made successfully in the past and you enjoyed vs. cooking up something new.
4. Reading a newspaper article from a journalist you like vs. reading something from a newcomer in the field.
5. A chimpanzee foraging in a new territory with unknown food resources as opposed to the known home territory.
Examples of exploration and exploitation in real life

1. Going to your favorite restaurant/bar vs. trying a new one.
2. Listening to music from a band you love vs. discovering new ones.
3. Preparing a meal that you have made successfully in the past and you enjoyed vs. cooking up something new.
4. Reading a newspaper article from a journalist you like vs. reading something from a newcomer in the field.
5. A chimpanzee foraging in a new territory with unknown food resources as opposed to the known home territory.
6. An organization trying a new organizational structure vs. a decently working existing one.
Multi-armed bandit (MAB) problem

Option 1

\[ N(\mu_1, \sigma_1) \]
\[ N(12, 3) \]

Option 2

\[ N(\mu_2, \sigma_2) \]
\[ N(15, 3) \]
History of the problem

1. The [MAB] problem was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied scientists that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage. — Whittle (1980)
History of the problem

1. The [MAB] problem was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied scientists that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage. — Whittle (1980)

2. The first papers and strategies on the topic were written by Thompson (1933) and Robins (1952).
History of the problem

The [MAB] problem was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied scientists that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage. — Whittle (1980)

The first papers and strategies on the topic were written by Thompson (1933) and Robins (1952).

Bellman and Gittins provided backward looking and forward looking solutions to the problem.
History of the problem

1. The [MAB] problem was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied scientists that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage. — Whittle (1980)

2. The first papers and strategies on the topic were written by Thompson (1933) and Robins (1952).

3. Bellman and Gittins provided backward looking and forward looking solutions to the problem.

4. Today the MAB framework is behind numerous algorithms that are used in the online world.

5. Note the similarities to the search problem considered last week: the problems fold into each other
Domains where MABs have been applied

1. Developing new medicine—clinical trials.
Domains where MABs have been applied

1. Developing new medicine—clinical trials.
2. One of the steam-engines for studying human (and animal) learning.
Domains where MABs have been applied

1. Developing new medicine—clinical trials.
2. One of the steam-engines for studying human (and animal) learning.
3. A very general framework for autonomous AI decision making. It is used as an alternative to A/B testing.
Domains where MABs have been applied

1. Developing new medicine—clinical trials.
2. One of the steam-engines for studying human (and animal) learning.
3. A very general framework for autonomous AI decision making. It is used as an alternative to A/B testing.
4. Currently used to allocate ads on the web. Companies like Criteo rely heavily on this framework.
Domains where MABs have been applied

1. Developing new medicine—clinical trials.
2. One of the steam-engines for studying human (and animal) learning.
3. A very general framework for autonomous AI decision making. It is used as an alternative to A/B testing.
4. Currently used to allocate ads on the web. Companies like Criteo rely heavily on this framework.
5. Used to decide which learning algorithm to use in a specific context.
Domains where MABs have been applied

1. Developing new medicine—clinical trials.
2. One of the steam-engines for studying human (and animal) learning.
3. A very general framework for autonomous AI decision making. It is used as an alternative to A/B testing.
4. Currently used to allocate ads on the web. Companies like Criteo rely heavily on this framework.
5. Used to decide which learning algorithm to use in a specific context.
6. Used to model how companies might choose among organizational structures or technologies of unknown merit.
Different strategies for coping with the multi-armed bandit problem

- Go optimal — not always possible and often computational very expensive.
Different strategies for coping with the multi-armed bandit problem

- Go optimal — not always possible and often computational very expensive.
- Go greedy — always try the best alternative.
Different strategies for coping with the multi-armed bandit problem

- Go optimal — not always possible and often computational very expensive.
- Go greedy — always try the best alternative.
- Add some noise — randomize one in while (e-greedy).
Different strategies for coping with the multi-armed bandit problem

- Go optimal — not always possible and often computational very expensive.
- Go greedy — always try the best alternative.
- Add some noise — randomize one in while (e-greedy).
- When randomizing choose options with higher expected return with a higher probability (softmax).
Different strategies for coping with the multi-armed bandit problem

- Go optimal — not always possible and often computational very expensive.
- Go greedy — always try the best alternative.
- Add some noise — randomize one in while (e-greedy).
- When randomizing choose options with higher expected return with a higher probability (softmax).
- Probability matching — choose actions according to their probability of being the best.
Different strategies for coping with the multi-armed bandit problem

- Go optimal — not always possible and often computational very expensive.
- Go greedy — always try the best alternative.
- Add some noise — randomize one in while (e-greedy).
- When randomizing choose options with higher expected return with a higher probability (softmax).
- Probability matching — choose actions according to their probability of being the best.
- Optimism in the face of uncertainty — prefer actions that are more uncertain, as they may turn out being rely good.
The Gittins index (Christian and Griffiths cpt. 2)

Possible to calculate for Bernoulli bandits with stable discounting of future trials.
A simple example (Sutton and Barto, cpt. 2)

Figure 2.1: An example bandit problem from the 10-armed testbed. The true value $q_*(a)$ of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean $q_*(a)$ unit variance normal distribution, as suggested by these gray distributions.
Performance of the $\varepsilon$-greedy algorithm

Figure 2.2: Average performance of $\varepsilon$-greedy action-value methods on the 10-armed testbed. These data are averages over 2000 runs with different bandit problems. All methods used sample averages as their action-value estimates.
Starting optimistically

Figure 2.3: The effect of optimistic initial action-value estimates on the 10-armed testbed. Both methods use a constant step-size parameter, $\alpha = 0.1$. 
Discussion: A/B testing and exploration-exploitation

The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions
The softmax rule

- Biases exploration towards the more promising actions.
The softmax rule

- Biases exploration towards the more promising actions.
- The softmax rule grades probabilities according to their selected values.
The softmax rule

- Biases exploration towards the more promising actions.
- The softmax rule grades probabilities according to their selected values.

\[
P(C(t) = j) = \frac{\exp(\theta E_j(t))}{\sum_{k=1}^{K} \exp(\theta E_k(t))}
\]
The softmax rule

- Biases exploration towards the more promising actions.
- The softmax rule grades probabilities according to their selected values.

\[
P(C(t) = j) = \frac{\exp(\theta E_j(t))}{\sum_{k=1}^{K} \exp(\theta E_k(t))}
\]

- where \( \theta \) is a temperature controlling how biased the algorithm will be.
Optimism in the face of uncertainty and the upper confidence bound (UCB)

- The more uncertain you are about the value of an option the more important it is to explore.
- That option could turn out to be really good and in the long-term improve your overall utility.
- UCB: \( P(C = i) \propto \exp(\theta m_i + \alpha \sqrt{\text{var}_i}) \)
The exploration-exploitation trade-off

Pantelis Pipergias Analytis

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions
Probability matching, changing environments and Thompson sampling

- Probability matching suggests sampling alternatives according to their rewards or their probability of being the best.
- Thompson sampling is an implementation of the probability matching principle.
Collective exploration

- The Roger’s paradox — produce or scrounge?
Collective exploration

- The Roger’s paradox — produce or scrounge?
- The social learning tournament — Rendell et al. (2010)
Collective exploration

- The Roger’s paradox — produce or scrounge?
- The social learning tournament — Rendell et al. (2010)
- Counter-intuitive more or less effects — Toyokawa et al. (2014)
Collective exploration

- The Roger’s paradox — produce or scrounge?
- The social learning tournament — Rendell et al. (2010)
- Counter-intuitive more or less effects — Toyokawa et al. (2014)
The typical bandit setting is like blind tasting...
My grandma’s problem: Choosing the best place to swim
The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

The machine learner’s problem

A contextual bandit experiment

Figure 1. Screenshot of the CMAB task in Experiment 1.
A contextual bandit experiment: Results

Figure 2. Results of the the continuous-linear CMAB task of Experiment 1. (a) average mean score per round, (b) proportion of choices of the 4th arm, and (c) proportion of choices of the best arm. Red error bars indicate standard error aggregated over 5 trials. Regression line is based on a least square regression including a 95% confidence level interval of the prediction line.
Contextual multi-armed bandit (CMAB) problem

Option 1

\[ N(f(\cdot), \sigma_1) \]
\[ N(w_1 x_1 + w_2 x_2, \sigma) \]
\[ N(\mu_1, \sigma_1) \]

Option 2

\[ N(f(\cdot), \sigma_2) \]
\[ N(w_1 x_1 + w_2 x_2, \sigma) \]
\[ N(\mu_2, \sigma_2) \]
Realistic decision problem...
The exploration-exploitation trade-off

Pantelis Pipergias
Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

Realistic decision problem...

Total number of rounds: 100
Current round: 1

Running total: 20
Motivation

Why is the CMAB problem interesting?

1. Captures the important characteristics of decisions in the wild better.
2. We can study how function learning interacts with decision making, how people deal with novelty, transfer of learning.
3. TD (\(\lambda\)) & curse of dimensionality - function learning as a solution. These problems are notoriously hard to solve using optimization techniques.
4. There is no realistic framework within we can study how people their preferences. CMAB might provide us with one.
Motivation

Why is the CMAB problem interesting?

1. Captures the important characteristics of decisions in the wild better.
Motivation

Why is the CMAB problem interesting?

1. Captures the important characteristics of decisions in the wild better.

2. We can study how function learning interacts with decision making, how people deal with novelty, transfer of learning.
Motivation

Why is the CMAB problem interesting?

1. Captures the important characteristics of decisions in the wild better.

2. We can study how function learning interacts with decision making, how people deal with novelty, transfer of learning.

3. $TD(\lambda)$ & curse of dimensionality - function learning as a solution. These problems are notoriously hard to solve using optimization techniques.
Motivation

Why is the CMAB problem interesting?

1. Captures the important characteristics of decisions in the wild better.

2. We can study how function learning interacts with decision making, how people deal with novelty, transfer of learning.

3. $TD(\lambda)$ & curse of dimensionality - function learning as a solution. These problems are notoriously hard to solve using optimization techniques.

4. There is no realistic framework within we can study how people their preferences. CMAB might provide us with one.
CMAB task

Total number of rounds: 100
Current round: 1
Running total: 20

Option 1

Option 2

Option 3

Option 4

Option 5

Option 6

Option 7

Option 8

Option 9

Option 10

Option 11

Option 12

Option 13

Option 14

Option 15

Option 16

Option 17

Option 18

Option 19

Option 20

Click on a square to choose an option. Press ENTER to continue to the next round.
The exploration-exploitation trade-off

Pantelis Pipergias Analytis

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

MAB task

Total number of rounds: 100
Current round: 1

Running total: 20

Option 1
Option 2
Option 3
Option 4
Option 5

Option 6
Option 7
Option 8
Option 9
Option 10

Option 11
Option 12
Option 13
Option 14
Option 15

Option 16
Option 17
Option 18
Option 19
Option 20

Click on a square to choose an option. Press ENTER to continue to the next round.
Three alternatives:

- **Dominating** - highest function value.
- **Neutral** - middle function value.
- **Dominated** - lowest function value.

Total number of rounds: 70
Current round: 5

Click on a square to choose an option. Press ENTER to continue to the next round.
Experimental Design

Training phase

- Between subject design – CMAB or MAB
- Contextual multi-armed bandit (CMAB) task – two informative features are visually displayed
- Classic multi-armed bandit (MAB) task – control group, features are not visible
- 20 alternatives, 100 trials

Test phase

- Designed to test the functional knowledge.
- One shot choices, no outcome feedback.
- 3 arms in 70 trials.
Gaussian process (GP) based “optimal” solutions

**Goal:** simultaneously **learn** and **optimize** unknown function.

\[ y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \]
**Goal:** simultaneously **learn** and **optimize** unknown function.

\[ y = f(x) + \epsilon, \ \epsilon \sim N(0, \sigma^2) \]

**GP based function learning process**

- \[ f(x) \sim GP(m(x), K(x, x')) \]
- \[ K(x, x') = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right) \]
Gaussian process (GP) based “optimal” solutions

**Goal:** simultaneously **learn** and **optimize** unknown function.

\[ y = f(x) + \epsilon, \  \epsilon \sim N(0, \sigma^2) \]

GP based function learning process

- \( f(x) \sim GP(m(x), K(x, x')) \)
- \( K(x, x') = \sigma_f^2 \exp(-\frac{(x-x')^2}{2l^2}) \)

Two versions of the choice process

1. **Upper confidence bound (GP-UCB):** \( \arg\max_i m_i + 2\sqrt{\text{var}_i} \)
2. **Thompson sampling (GP-Th):** Draw from \( p(\theta|D, M) \) for each arm, take the max.
The exploration-exploitation trade-off

Pantelis Pipergias
Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

GP prior, 1D example
GP-Thompson, 1D example

The exploration-exploitation trade-off

Pantelis Pipergias
Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions
How much do people rely on knowledge of the relationships between features and alternative value’s when making decisions?

Can we model people’s behavior using traditional machine learning models?

How priors about functional relationships affect the decision making?

Do people explore the choice set strategically, to learn the relationships?
Experiment 1 – Positive linear function

Experiment 1a – Amazon Turk

- Feature values $x$ drawn from $U(0.1, 0.9)$
- For each arm $j$ in trial $t$, the payoffs $R_j(t)$ were computed as:

$$R_j(t) = 2 \times x_{1,j} + 1 \times x_{2,j} + \epsilon_j(t).$$

- $\epsilon_j(t)$ drawn independently for each arm in every trial, from $N(0, 0.25)$.
- Task was to maximize the cumulative reward.
- 186 participants – monetary payoffs.
Experiment 1 – Positive linear function

Experiment 1a – Amazon Turk
- Feature values $x$ drawn from $U(0.1, 0.9)$
- For each arm $j$ in trial $t$, the payoffs $R_j(t)$ were computed as:

$$R_j(t) = 2 \times x_{1,j} + 1 \times x_{2,j} + \epsilon_j(t).$$

- $\epsilon_j(t)$ drawn independently for each arm in every trial, from $N(0, 0.25)$.
- Task was to maximize the cumulative reward.
- 186 participants – monetary payoffs.

Experiment 1b – lab replication
- Weights and noise rescaled: $w_1 = 20, w_2 = 10, N(0, 2.5)$.
- 75 UPF lab participants – monetary payoffs.
The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Exploration-exploitation problems
The multi-armed bandit framework
Strategies
Contextual bandits
Results from a real world experiment
Conclusions

Mean choice rank - Exp 1a

Mean rank of the chosen alternative

Random performance

Block

Mean rank of the chosen alternative

MAB
CMAB
GP−UCB

Individual 1
Individual 2
The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

Mean choice rank - Exp 1a

Mean rank of the chosen alternative

MAB
CMAB
GP−UCB

Random performance

Block

1 2 3 4 5

1 2 3 4 5

Mean rank of the chosen alternative

1 2 3 4 5

1 2 3 4 5

1 2 3 4 5

Mean choice rank - Exp 1b
One-shot choices in the test phase

Three alternatives:

- **Dominating** - highest function value.
- **Neutral** - middle function value.
- **Dominated** - lowest function value.

Total number of rounds: 70
Current round: 5

Click on a square to choose an option. Press ENTER to continue to the next round.
One-shot choices in the test phase

- CMABn Diff/Extra
- CMABn Diff/Inter
- CMABn Easy/Extra
- CMABn Easy/Inter
- CMABn Weight test

Mean proportion of choices

Rank of the chosen alternative

Results from a real world experiment

Conclusions

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Pantelis Pipergias Analytis

The exploration-exploitation trade-off
One-shot choices in the test phase

The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

One-shot choices – Lab replication
Exploration in the feature space

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions
Exploration in the feature space – First 10 trials

The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

Lab replication

Clusters: Exploration
Exploration in the feature space – All trials

MAB

CMABn

Feature 1

Feature 2

(0.1,0.3]

(0.3,0.5]

(0.5,0.7]

(0.7,0.9]

Proportion

0.3

0.2

0.1

Exp 1b - Lab replication
Inter-individual differences: Function-based and naive learners
Clustering according to the test phase performance

The exploration-exploitation trade-off

Pantelis Pipergias

Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions
Clusters: Performance in the CMAB task

N1 = 43
N2 = 53
R_{tr} = 7
R_{tr} = 4.59
R_{te} = 1.94
R_{te} = 1.24
Clustering: Feature space, first 10 trials

- Feature 1
  - (0.1, 0.3)
  - (0.3, 0.5)
  - (0.5, 0.7)
  - (0.7, 0.9)

- Feature 2
  - (0.1, 0.3)
  - (0.3, 0.5)
  - (0.5, 0.7)
  - (0.7, 0.9)

Proportion

- CMABn 1
- CMABn 2

- Exploration in the MAB condition
How much do people rely on knowledge of the relationships between features and alternative value’s when making decisions?

Can we model people’s behavior using traditional machine learning models?

How priors about functional relationships affect the decision making?

Do people explore the choice set strategically, to learn the relationships?
Modeling user behavior

- **Learning**: We model participants as function learners (GP) or as tracing mean rewards (BMT)
  1. Gaussian processes (GP) function learning model:
     \[ f(x) \sim GP(m(x), K(x, x')), K(x, x') = \sigma_f^2 \exp(-\frac{(x-x')^2}{2l^2}) \]
  2. Bayesian mean reward tracing (BMT)

- **Choices**: Participants either use uncertainty in balancing the exploration-exploitation (UCB) or not (SM).
  1. Upper confidence bound (UCB):
     \[ P(C = i) \propto \exp(\theta m_i + \alpha \sqrt{\text{var}_i}) \]
  2. Softmax (SM): \[ P(C = i) \propto \exp(\theta m_i) \]
There is evidence for GP models, especially for participants that know function well (according to the test task). Models with UCB perform poorly.
How much do people rely on knowledge of the relationships between features and alternative value’s when making decisions?

Can we model people’s behavior using traditional machine learning models?

How priors about functional relationships affect the decision making?

Do people explore the choice set strategically, to learn the relationships?
Experiment 1c – Quadratic and mixed linear function

Training phase

- 2x2 between subject design: Type of task (CMAB, MAB) and Type of function (Quadratic, Mixed)
- Quadratic function:
  \[1 + 60(x_1 - .02)^2 + 60(x_1 - .02)^2 + 30x_1x_2, \ N(0, 2.5)\]
- Mixed linear function: \(w_1 = 40, \ w_2 = -30, \ N(0, 2.5)\)
- 376 participants – Amazon Turk – monetary payoffs.

Test phase

- Test items for mixed linear function are the same as for the positive linear one
- Special items for the quadratic function, testing whether people detected the nonlinear nature of the relationship.
Exploration in the feature space, first 10 trials

MAB mixed

CMAB mixed

MAB quadratic

CMAB quadratic

Feature 2

Feature 1

Proportion

0.100

0.075

0.050
Exploration in the feature space, all trials

The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions
How much do people rely on knowledge of the relationships between features and alternative value’s when making decisions?

Can we model people’s behavior using traditional machine learning models?

How priors about functional relationships affect the decision making?

Do people explore the choice set strategically, to learn the relationships?
Experiment 2 – Function learning pretraining

Exploration to learn the function should depend on…
Experiment 2 – Function learning pretraining

Exploration to learn the function should depend on...
- Uncertainty about the function.
Experiment 2 – Function learning pretraining

Exploration to learn the function should depend on...

- Uncertainty about the function.
- Type of function.
Experiment 2 – Function learning pretraining

Exploration to learn the function should depend on...
- Uncertainty about the function.
- Type of function.
- Horizon.
Experiment 2 – Function learning pretraining

Exploration to learn the function should depend on...
- Uncertainty about the function.
- Type of function.
- Horizon.
- Expecting need for generalization.
Experiment 2 – Function learning pretraining

Exploration to learn the function should depend on…
- Uncertainty about the function.
- Type of function.
- Horizon.
- Expecting need for generalization.

Training phase
- Mixed design – Two between factors: Type of function (Positive linear, Quadratic) x Horizon (100 or 30 trials CMAB phase) and within factor (With or without function learning phase)
Experiment 2 – Function learning pretraining

Exploration to learn the function should depend on...
- Uncertainty about the function.
- Type of function.
- Horizon.
- Expecting need for generalization.

**Training phase**
- Mixed design – Two between factors: Type of function (Positive linear, Quadratic) x Horizon (100 or 30 trials CMAB phase) and within factor (With or without function learning phase)
- Function learning task – 100 trials with single alternative, same two features and function, accuracy incentivized.
Experiment 2 – Function learning pretraining

Exploration to learn the function should depend on...

- Uncertainty about the function.
- Type of function.
- Horizon.
- Expecting need for generalization.

Training phase

- Mixed design – Two between factors: Type of function (Positive linear, Quadratic) x Horizon (100 or 30 trials CMAB phase) and within factor (With or without function learning phase)
- Function learning task – 100 trials with single alternative, same two features and function, accuracy incentivized.
- Same positive linear and quadratic functions as before, but alternatives now includes randomly drawn intercepts!
- 425 participants – Amazon Turk – monetary payoffs.
Mean choice ranks

The exploration-exploitation trade-off
Pantelis Pipergias Analytis

Exploration-exploitation problems
The multi-armed bandit framework
Strategies
Contextual bandits
Results from a real world experiment
Conclusions

Mean rank of the chosen alternative
Random performance

Block

Linear

Quadratic

CMAB linear
fCMAB linear
fCMABs linear
CMAB quadratic
fCMAB quadratic
fCMABs quadratic
Exploration in the feature space, first 10 trials

CMAB lin

CMAB quad

fCMAB lin

fCMAB quad

fCMABs lin

fCMABs quad
### Exploration in the feature space, all trials

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Feature 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.1,0.3)</td>
<td>(0.1,0.3)</td>
</tr>
<tr>
<td>(0.3,0.5)</td>
<td>(0.3,0.5)</td>
</tr>
<tr>
<td>(0.5,0.7)</td>
<td>(0.5,0.7)</td>
</tr>
<tr>
<td>(0.7,0.9)</td>
<td>(0.7,0.9)</td>
</tr>
</tbody>
</table>

#### CMAB lin

- (0.1,0.3)
- (0.3,0.5)
- (0.5,0.7)
- (0.7,0.9)

#### fCMAB lin

- (0.1,0.3)
- (0.3,0.5)
- (0.5,0.7)
- (0.7,0.9)

#### fCMABs lin

- (0.1,0.3)
- (0.3,0.5)
- (0.5,0.7)
- (0.7,0.9)

#### CMAB quad

- (0.1,0.3)
- (0.3,0.5)
- (0.5,0.7)
- (0.7,0.9)

#### fCMAB quad

- (0.1,0.3)
- (0.3,0.5)
- (0.5,0.7)
- (0.7,0.9)

#### fCMABs quad

- (0.1,0.3)
- (0.3,0.5)
- (0.5,0.7)
- (0.7,0.9)
Summary

- People learn the function and generalize their knowledge to new decision situations.
Summary

■ People learn the function and generalize their knowledge to new decision situations.

■ But there are inter-individual differences – some people rely on learning the function, others are naive learners; akin to model-based vs model-free RL.
Summary

- People learn the function and generalize their knowledge to new decision situations.
- But there are inter-individual differences – some people rely on learning the function, others are naive learners; akin to model-based vs model-free RL.
- New flavour of the exploration-exploitation trade-off – evidence that people simultaneously learn and optimize the function.
People learn the function and generalize their knowledge to new decision situations.

But there are inter-individual differences – some people rely on learning the function, others are naive learners; akin to model-based vs model-free RL.

New flavour of the exploration-exploitation trade-off – evidence that people simultaneously learn and optimize the function.

Priors about the functional relationship can hurt the performance.
People learn the function and generalize their knowledge to new decision situations.

But there are inter-individual differences – some people rely on learning the function, others are naive learners; akin to model-based vs model-free RL.

New flavour of the exploration-exploitation trade-off – evidence that people simultaneously learn and optimize the function.

Priors about the functional relationship can hurt the performance.

People do not seem to take into account the time horizon. People exploit more aggressively when they have been pre-trained on the function.
Challenges and future directions

- Goals is to develop a function learning based RL model – algorithmic level.
Challenges and future directions

- Goals is to develop a function learning based RL model – algorithmic level.
- Moreover, it is difficult to fit function learning models without prediction data. However, asking for predictions along with choices changes the behaviour.
Challenges and future directions

- Goals is to develop a function learning based RL model – algorithmic level.
- Moreover, it is difficult to fit function learning models without prediction data. However, asking for predictions along with choices changes the behaviour.
- How do people behave in the presence of information about the alternatives and other contextual information?
Challenges and future directions

- Goals is to develop a function learning based RL model – algorithmic level.
- Moreover, it is difficult to fit function learning models without prediction data. However, asking for predictions along with choices changes the behaviour.
- How do people behave in the presence of information about the alternatives and other contextual information?
Acknowledgments

Funding:

- FPU grant, Ministry of Education, Culture and Sports, Spain
- Max Planck Institute for Human Development, Berlin
- Barcelona Graduate School of Economics
Quadratic function – An illustration
Individual behaviour in the training phase - Experiment 1a

Choice behavior of subject e2–0124, CMABn condition, experiment LowNoise

Rank of the chosen alternative

Mean choice rank

0 25 50 75 100
Individual behaviour in the training phase - Experiment 1a

Choice behavior of subject e2–0065, CMABn condition, experiment LowNoise

Rank of the chosen alternative

Mean choice rank
The exploration-exploitation trade-off

Pantelis Pipergias

Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

Mean choice rank – Lab replication

![Graph showing mean rank of the chosen alternative over blocks for MAB and CMAB strategies. The graph compares the performance of MAB and CMAB with random performance. The y-axis represents the mean rank of the chosen alternative, and the x-axis represents the block number. The graph shows a downward trend for both MAB and CMAB, indicating improved performance as the experiment progresses. The random performance line remains constant.](image-url)
The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

One-shot choices – Lab replication

Mean proportion of choices

Rank of the chosen alternative

One-shot choices - Exp 1a
Feature space, all trials – Lab replication

The exploration-exploitation trade-off
Pantelis Pipergias Analytis

Exploration-exploitation problems
The multi-armed bandit framework
Strategies
Contextual bandits
Results from a real world experiment
Conclusions

Feature 1
(0.1,0.3] (0.3,0.5] (0.5,0.7] (0.7,0.9]

Feature 2
(0.1,0.3] (0.3,0.5] (0.5,0.7] (0.7,0.9]

MAB
CMABn

Proportion
0.1 0.2 0.3

Back to Exp 1a
Feature space, first 10 trials – Lab replication

MAB

CMABn

Feature 1

Feature 2

(0.1,0.3)
(0.3,0.5)
(0.5,0.7)
(0.7,0.9)

(0.1,0.3)
(0.3,0.5)
(0.5,0.7)
(0.7,0.9)

Proportion

0.25
0.20
0.15
0.10
0.05
Mean choice rank – Mixed and quadratic

<table>
<thead>
<tr>
<th>Block</th>
<th>Mixed</th>
<th>Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Mixed**
  - MAB mixed
  - CMAB mixed
  - Random performance

- **Quadratic**
  - MAB quadratic
  - CMAB quadratic
  - Random performance

*Feature space, all* ✔ *Feature space, subset* ✔
One-shot choices – Mixed and quadratic

The exploration-exploitation trade-off

Pantelis Pipergias Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions
Mean choice rank – Positive and quadratic

**Linear**

- Random performance
- Mean rank of the chosen alternative

**Quadratic**

- Random performance
- Mean rank of the chosen alternative

**Results from a real world experiment**

- CMAB linear
- fCMAB linear
- fCMABs linear
- CMAB quadratic
- fCMAB quadratic
- fCMABs quadratic

**Strategies**

- Feature space, all
- Feature space, subset
One-shot choices – Positive linear

CMAB lin
Easy

CMAB lin
Difficult

CMAB lin
Weight test

fCMAB lin
Easy

fCMAB lin
Difficult

fCMAB lin
Weight test

Mean proportion of choices

Rank of the chosen alternative

Feature space, all
Feature space, subset

73 / 75
The exploration-exploitation trade-off

Pantelis Pipergias

Analytis

Exploration-exploitation problems

The multi-armed bandit framework

Strategies

Contextual bandits

Results from a real world experiment

Conclusions

One-shot choices – Quadratic

![Graph showing the mean proportion of choices for different tests and strategies.](image-url)

- CMAB q
  - Max test 1
  - Min test 1
  - Slope test 1
  - Max test 2
  - Min test 2
  - Slope test 2

- fCMAB q
  - Max test 1
  - Min test 1
  - Slope test 1
  - Max test 2
  - Min test 2
  - Slope test 2

Rank of the chosen alternative

Mean proportion of choices

Feature space, all

Feature space, subset