Personalized Preference Learning
from Spinal Cord Stimulation to Exoskeletons

Yisong Yue
Goal-Oriented Experiment Design
(aka "Interactive Learning")

- Iterative & adaptive
- Utility maximizing (for a suitable definition of utility)
Personalized Therapies

Medtronic human array

Each patient is unique
10⁹ possible configurations!

Image source: williamcapicottomd.com

SCI Patient

Yanan Sui
Personalized Therapies

Atalante Exoskeleton by Wandercraft
Key Question: How to Elicit Feedback?

Absolute Feedback: “That felt good, 4/5 rating.”

Challenge: humans are not consistent in providing absolute feedback.
Our Approach: Preference Feedback

Multi-dueling Bandits with Dependent Arms, Sui, Zhuang, Burdick & Yue, UAI 2017
Correlational Dueling Bandits with Application to Clinical Treatment in Large Decision Spaces, Sui, Yue & Burdick, IJCAI 2017
Preference-Based Learning for Exoskeleton Gait Optimization, Tucker, Novoseller, et al., ICRA 2020  (Best Paper Award)
Human Preference-Based Learning for High-dimensional Optimization of Exoskeleton Walking Gaits, Tucker et al., IROS 2020
ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes, Li, Tucker, et al., ICRA 2021
Preference-Based Interactive Learning

Core Algorithms & Theory

The $K$-armed Dueling Bandits Problem

Bayesian Active Learning for Classification and Preference Learning

Towards Preference-Based Reinforcement Learning

Applications to Computer Systems

Preference Elicitation for Interface Optimization

Active Exploration for Learning Rankings from Clickthrough Data

Applications to Autonomous Systems

Learning Preferences for Manipulation Tasks from Online Coactive Feedback

Active Preference-Based Learning of Reward Functions
Remainder of Talk

Sparring Family of Algorithms

• Sparring  [Ailon, Karnin, Joachims, ICML 2014]
• SelfSparring  [Sui, Zhuang, Burdick, Yue, UAI 2017]
• CoSpar  [Tucker, Novoseller, Kann, Sui, Yue, Burdick, Ames, ICRA 2020]  (Best Paper Award)
• LineCoSpar  [Tucker, Cheng, Novoseller, Cheng, Yue, Burdick, Ames, IROS 2020]

Briefly Survey Other Considerations
(Basic) Interaction Protocol

1. Choose two actions \((x_1, x_2)\)
2. Compare \(x_1\) vs \(x_2\) \((stochastic)\)
3. Feedback \(y=1\) if \(x_1\) wins \((y=0\ o.w.)\)
4. Repeat

Example Preference Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.03</td>
<td>0.04</td>
<td>0.06</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>B</td>
<td>-0.03</td>
<td>0</td>
<td>0.03</td>
<td>0.05</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>C</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0</td>
<td>0.04</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>D</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.04</td>
<td>0</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>E</td>
<td>-0.10</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.05</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>F</td>
<td>-0.11</td>
<td>-0.11</td>
<td>-0.09</td>
<td>-0.07</td>
<td>-0.03</td>
<td>0</td>
</tr>
</tbody>
</table>

Values are \(Pr(\text{row} > \text{col}) – 0.5\)

Measurements are relative rather than absolute!
(Basic) Interaction Protocol

1. Choose two actions \((x_1, x_2)\)
2. Compare \(x_1\) vs \(x_2\) \((stochastic)\)
3. Feedback \(y=1\) if \(x_1\) wins \((y=0\ o.w.)\)
4. Repeat

Example Preference Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00</td>
<td>0.03</td>
<td>0.04</td>
<td>0.06</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>B</td>
<td>-0.03</td>
<td>0.00</td>
<td>0.03</td>
<td>0.05</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>C</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0.00</td>
<td>0.04</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>D</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.04</td>
<td>0.00</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>E</td>
<td>-0.10</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td>F</td>
<td>-0.11</td>
<td>-0.11</td>
<td>-0.09</td>
<td>-0.07</td>
<td>-0.03</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Values are \(\text{Pr}(\text{row} > \text{col}) - 0.5\)

The K-armed Dueling Bandits Problem, Yue, Broder, Kleinberg, Joachims, COLT 2009

Measurements are relative rather than absolute!
(Basic) Interaction Protocol

1. Choose two actions \((x_1, x_2)\)
2. Compare \(x_1\) vs \(x_2\) (stochastic)
3. Feedback \(y=1\) if \(x_1\) wins \((y=0\) o.w.\)
4. Repeat

Comparing \(A\) & \(C\): \(P(A > C) = 0.54\)

Example Preference Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.03</td>
<td>0.04</td>
<td>0.06</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>B</td>
<td>-0.03</td>
<td>0</td>
<td>0.03</td>
<td>0.05</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>C</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0</td>
<td>0.04</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>D</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.04</td>
<td>0</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>E</td>
<td>-0.10</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.05</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>F</td>
<td>-0.11</td>
<td>-0.11</td>
<td>-0.09</td>
<td>-0.07</td>
<td>-0.03</td>
<td>0</td>
</tr>
</tbody>
</table>

Values are \(Pr(\text{row} > \text{col}) - 0.5\)

Measurements are relative rather than absolute!

The Karmed Dueling Bandits Problem, Yue, Broder, Kleinberg, Joachims, COLT 2009
(Basic) Interaction Protocol

1. Choose two actions \((x_1, x_2)\)
2. Compare \(x_1\) vs \(x_2\) (stochastic)
3. Feedback \(y=1\) if \(x_1\) wins \((y=0\) o.w.\)
4. Repeat

Active Learning: Identify Entire Matrix

Optimization/Bandits: Converge to Best Action
Our Main Focus!

Example Preference Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.03</td>
<td>0.04</td>
<td>0.06</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>B</td>
<td>-0.03</td>
<td>0</td>
<td>0.03</td>
<td>0.05</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>C</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0</td>
<td>0.04</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>D</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.04</td>
<td>0</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>E</td>
<td>-0.10</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.05</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>F</td>
<td>-0.11</td>
<td>-0.11</td>
<td>-0.09</td>
<td>-0.07</td>
<td>-0.03</td>
<td>0</td>
</tr>
</tbody>
</table>

Measurements are relative rather than absolute!

\(\text{Compare A & A: } P(A > A) = 0.50\)

Values are \(\text{Pr(row > col)} - 0.5\)

The K-armed Dueling Bandits Problem, Yue, Broder, Kleinberg, Joachims, COLT 2009
(Prior Work) Sparring

• Instantiate 2 Bandit algorithms: $P_1$ & $P_2$

• For $t = 1$, ...
  
  • $P_1$ chooses $x_1$
  
  • $P_2$ chooses $x_2$
  
  • Compare $x_1$ vs $x_2$
  
  • Provide feedback
    
    • $P_1$ observes if $x_1$ won
    
    • $P_2$ observes if $x_2$ won

Reducing Dueling Bandits to Cardinal Bandits, Ailon, Karnin & Joachims, ICML 2014
Intuition

• Reduction to standard Bandit settings
  – Each player selfishly maximizes own reward

  • Instantiate $P_1$
  • For $t = 1, \ldots$
    • $P_1$ chooses $x_1$
    • Plays $x_1$
    • Observes feedback

  • Instantiate $P_2$
  • For $t = 1, \ldots$
    • $P_2$ chooses $x_2$
    • Plays $x_2$
    • Observes feedback

Challenges:
• Hard to analyze
• Inefficient (throws away information)
SelfSparring

- Instantiate 1 Bandit algorithm P
- For $t = 1, \ldots$
  - P chooses $x_1$
  - P chooses $x_2$
  - Compare $x_1$ vs $x_2$
  - Provide feedback
    - P observes both sides

Probabilistic Bandit Algorithm (Thompson Sampling)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.03</td>
<td>0.04</td>
<td>0.06</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>B</td>
<td>-0.03</td>
<td>0</td>
<td>0.03</td>
<td>0.05</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>C</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0</td>
<td>0.04</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>D</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.04</td>
<td>0</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>E</td>
<td>-0.10</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.05</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>F</td>
<td>-0.11</td>
<td>-0.11</td>
<td>-0.09</td>
<td>-0.07</td>
<td>-0.03</td>
<td>0</td>
</tr>
</tbody>
</table>

Multi-dueling Bandits with Dependent Arms, Sui, Zhuang, Burdick & Yue, UAI 2017
Benefits of SelfSparring

• Rigorous convergence analysis (regret bound)

• Performs very well in practice

• Easily extendable to new settings

Multi-dueling Bandits with Dependent Arms, Sui, Zhuang, Burdick & Yue, UAI 2017
Extension: Comparing Multiple Actions

• Need to consider multiple actions before giving feedback

• E.g., rank a batch of 5 electro-stimulation stimuli

• SelfSparring trivially generalizes!

Multi-dueling Bandits with Dependent Arms, Sui, Zhuang, Burdick & Yue, UAI 2017
Clinical Experiments
Extension: Mixed-Initiative Feedback

Maegan Tucker
Ellen Novoseller

 preference-based learning for exoskeleton gait optimization, maegan tucker, ellen novoseller, et al., icra 2020 (best paper award)
Unsafe Actions

• Avoiding Unsafe or Undesirable Regions
• Requires Absolute Feedback from Users
  • “This feels uncomfortable / painful.”

Standard Bandits/BayesOpt Algorithms

Safe Region

Action Space

ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes, Amy Kejun Li et al., ICRA 2021
Stagewise Safe Bayesian Optimization with Gaussian Processes, Yanan Sui et al., ICML 2018
High Dimensions

LineCoSpar: Optimize series of 1-D subspaces (builds on LineBO alg.)

Human Preference-Based Learning for High-dimensional Optimization of Exoskeleton Walking Gaits, Maegan Tucker et al., IROS 2020
Optimization vs Active Learning

### Optimization (CoSpar)
- Reliable information only around the optimum
- Identify optimal gaits for specific users

### Active Learning (ROIAL)
- Reliable information for the entire landscape
- Better understand the science of walking
- More likely to give users “bad” actions

**ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes**, Amy Kejun Li et al., ICRA 2021

[Images and text extracts related to optimization and active learning, along with the authors' names Amy Li, Maegan Tucker, Erdem Biyik.]
The mission of the Robotic Assisted Mobility Science (RoAMS) initiative is to bring together research frontiers across multiple labs to restore mobility to individuals with a condition that limits or prevents stable locomotion. These research frontiers include mechanical design, machine learning, and formal control algorithm synthesis. This important research has the potential to offer advancements in robotic assistive devices such as prostheses, exoskeleton, exosuits, and spinal cord implants.

http://roams.caltech.edu/
Multi-dueling Bandits with Dependent Arms, Yanan Sui et al., UAI 2017
Correlational Dueling Bandits with Application to Clinical Treatment in Large Decision Spaces, Yanan Sui et al., IJCAI 2017
Preference-Based Learning for Exoskeleton Gait Optimization, Maegan Tucker, Ellen Novoseller et al., ICRA 2020 (Best Paper Award)
Human Preference-Based Learning for High-dimensional Optimization of Exoskeleton Walking Gaits, Maegan Tucker et al., IROS 2020
ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes, Amy Kejun Li et al., ICRA 2021
Stagewise Safe Bayesian Optimization with Gaussian Processes, Yanan Sui et al., ICML 2018