

Bandit Algorithms A Road Map!

Hunter Heidenreich

(Guided by Bandit Algorithms by Tor Lattimore and Csaba Szepesvári)

Questions to Ask

What does the problem action space look like?

Is there any structure to the problem?

Is there external information that my learner has access to?

How are rewards generated?

What kind of feedback does my learner receive each round?

Action Space

Infinite versus finite

- Finite, discretized
 - Ex. Select an arm in a 3-arm bandit
- Infinite, continuous
 - Ex. Select a number in the range (0, 1)

Single action versus combinatorial actions

- Single: Selection of 1 action
- Combinatorial: Selection of a vector of actions
 - Ex. Edges in a graph will be dropped with IID probabilities. Select a subset to form a path from node t to node s

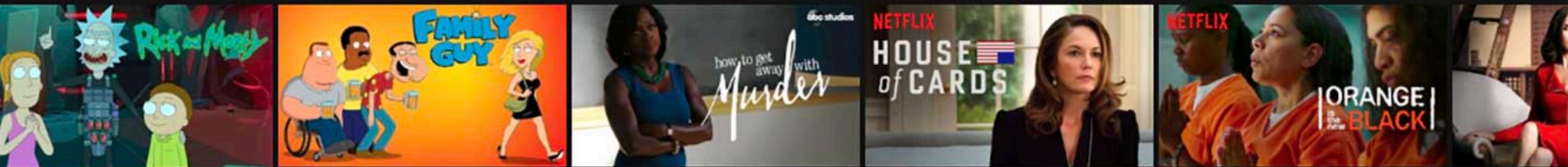
A



B



Emmy-winning US TV Shows



Police Detective TV Dramas



Critically Acclaimed Witty TV Shows



Problem Structure

- Unstructured
 - Selecting an action **does not** tell the learner anything about another action
 - Ex. A slot machine with IID rewards
- Structured
 - Selecting an action allows inference about what other actions may yield
 - Ex. A 2-armed Bernoulli bandit where arm 1 is parameterized by theta and arm 2 is parameterized by $(1 - \theta)$.
 - Ex. Linear Bandits



External Information

Is there a context?

- Contextual bandits (otherwise highly non-stationary)
- Ex. Selection of ads, given a vector of user information

Is there a prior?

- Bayesian bandits



Reward Mechanism

Stochastic

- Each action corresponds to an IID reward
- Mean rewards do not shift (significantly) over time

Non-stationary

- Reward distributions **do** shift, but given some sort of rule
- Relaxation of stochastic (at a cost)

Adversarial

- Adversary selects worst-case rewards given knowledge of learner's policy
- Randomization is key

Learner Feedback

Bandit Feedback	Partial Feedback (Semi-bandit feedback)	Full Feedback	Partial Monitoring
<ul style="list-style-type: none">• Only action selected (or total action reward in combinatorial setting)• Ex. Slot machine payout. No knowledge of what could've been gained from other arms	<ul style="list-style-type: none">• Combinatorial setting, each sub-action has associated reward• Ex. Learning which parts of the graph “dropped” in path finding	<ul style="list-style-type: none">• See all reward signals for every action• Useful in adversarial settings• Nonsensical in stochastic setting	<ul style="list-style-type: none">• Feedback is not received every round• Not a bandit problem

Common Formulations

Stochastic bandits
with finitely many
arms

- Unstructured, bandit feedback
- Part II of Bandit Algorithms

Adversarial bandits
with finitely many
arms

- Unstructured, bandit or full feedback
- Part III of Bandit Algorithms

Contextual and
Linear Bandits

- Structured!
- Part V and VI of Bandit Algorithms

Combinatorial,
Bayesian, and Non-
Stationary Bandits

- Other types of structure!
- Part VII of Bandit Algorithms

