

Making sequential decisions under uncertainty

You have N choices

In every round $t = 1, 2, \dots, T$


- Choose one out of N , using only past observations
- observe (uncertain) reward, feedback

Maximize total reward, other objective

Examples

- Movie Recommendation, online advertising
 - Observe clicks, likes
- Portfolio optimization
 - Observe money made (reward), how stocks behaved, market index changes etc.
- Pricing and Revenue management
 - Observe demand, revenue from sales
- Game playing
 - Observe improvement in the player position, other game state
- Robot navigation and control
 - Observe performance and accuracy

Common elements of sequential decision making models

- Can use only past feedback
 - Feedback from time steps before this
 - Past feedback has some relation to future reward
- 
- Learn from past to predict future and optimize

Distinctions

How past is related to future?

- Stochastic process
 - IID
 - Markovian
- Adversarial
 - An arbitrary sequence of feedback, but restricted in certain specific ways

Distinctions

- Full information models

- *Reward* for your decision but *feedback* on instantaneous performance of all possible choices, **Can answer what-if**
- E.g. buying stocks, pay-per-impression advertising, bidding/offer model of selling goods
- Online packing, online matching, online convex programming

- Limited feedback models

- E.g., Feedback only on performance of your decision
- E.g. movie recommendations, pay-per-click advertising, posted price model of selling goods, game playing
- Multi-armed bandits, Reinforcement learning

Managing exploitation-exploitation tradeoff

The **multi-armed bandit problem**
(Thompson 1933; Robbins 1952)

Multiple rigged slot machines in a casino.
Which one to put money on?

- Try each one out



***WHEN TO STOP TRYING (EXPLORATION) AND START PLAYING
(EXPLOITATION)?***

Multi-armed bandit model

- N arms (choices). Pulling an arm generates a reward
- In each round t , pull one arm I_t of the N arms.
- Observe stochastic reward r_t
- Maximize $\sum_t r_t$
- You do not know what you would get by pulling another arm
- Limited feedback or “bandit feedback”



Key properties of the model

- You observe the feedback for only the decision you make
 - WHAT IF had pulled another arm?
 - Need to explore
- Natural instinct is to take best choice according to current data
 - Exploitation
- Algorithms to manage the Exploration-Exploitation tradeoff
 - Adapt from learnings so far, to waste less time on exploring bad choices

Examples: Clinical trials

- Patients arrive sequentially
- Pick one out of N treatments
- Cure as many patients as possible
- You can observe the performance of a treatment by administration

Administer the currently best performing treatment, Or
Try a less understood treatment?



On the Likelihood that
One Unknown Probability
Exceeds Another in View
of the Evidence of Two
Samples

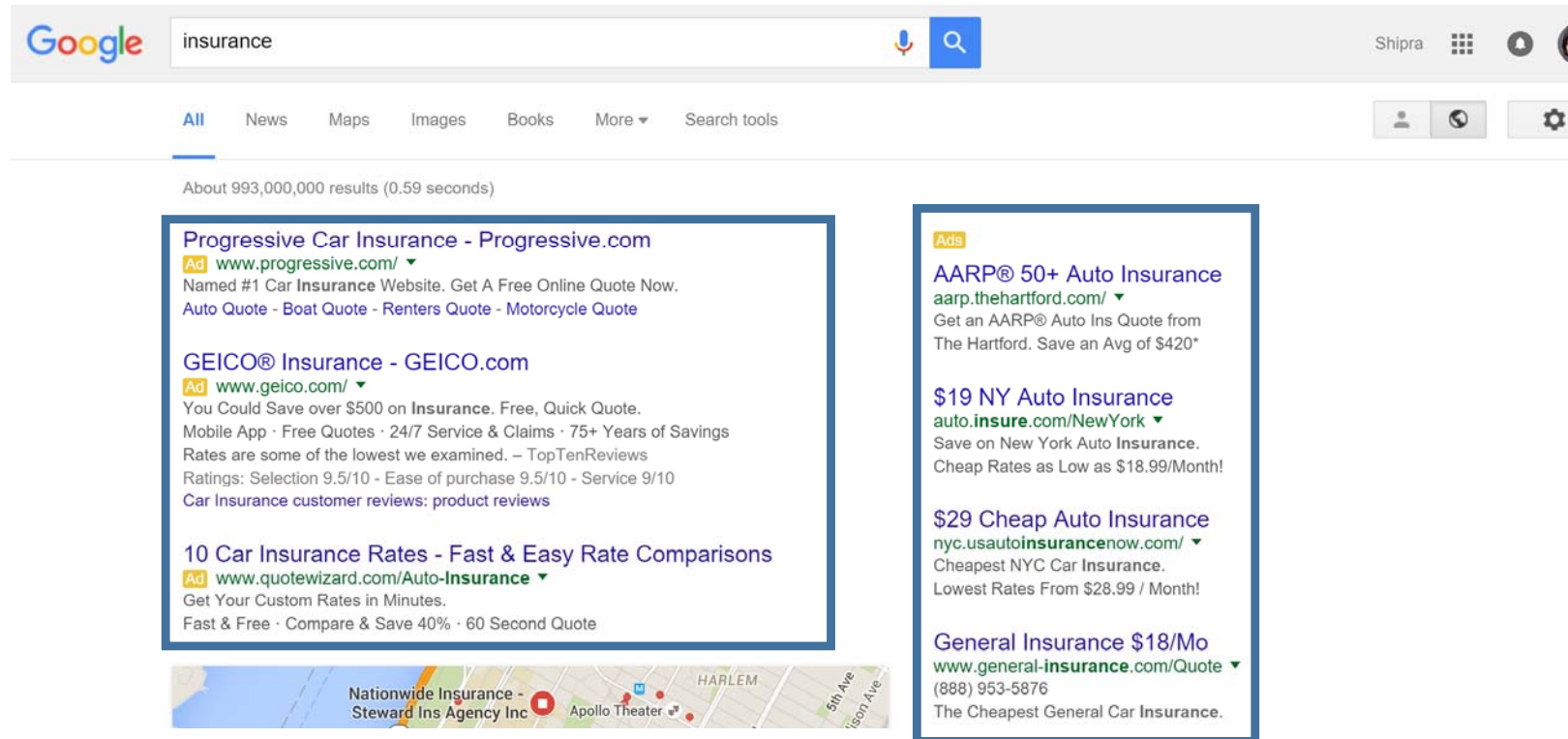
William R. Thompson

Biometrika

Vol. 25, No. 3/4 (Dec., 1933), pp.
285-294

Response depend on patient features, one patient informs about others: [Contextual bandits](#)

Internet advertising: pick a few from N ads



The screenshot shows a Google search for "insurance". The search bar at the top contains the word "insurance" and a microphone icon. Below the search bar, there are tabs for "All", "News", "Maps", "Images", "Books", "More", and "Search tools". The "All" tab is selected. Below the tabs, it says "About 993,000,000 results (0.59 seconds)".

There are two main columns of ads. The left column contains three ads:

- Progressive Car Insurance - Progressive.com**
Ad www.progressive.com/ ▼
Named #1 Car Insurance Website. Get A Free Online Quote Now.
Auto Quote - Boat Quote - Renters Quote - Motorcycle Quote
- GEICO® Insurance - GEICO.com**
Ad www.geico.com/ ▼
You Could Save over \$500 on Insurance. Free, Quick Quote.
Mobile App · Free Quotes · 24/7 Service & Claims · 75+ Years of Savings
Rates are some of the lowest we examined. – TopTenReviews
Ratings: Selection 9.5/10 - Ease of purchase 9.5/10 - Service 9/10
Car Insurance customer reviews: product reviews
- 10 Car Insurance Rates - Fast & Easy Rate Comparisons**
Ad www.quotewizard.com/Auto-Insurance ▼
Get Your Custom Rates in Minutes.
Fast & Free · Compare & Save 40% · 60 Second Quote

The right column contains three ads:

- AARP® 50+ Auto Insurance**
Ad aarp.thehartford.com/ ▼
Get an AARP® Auto Ins Quote from The Hartford. Save an Avg of \$420*
- \$19 NY Auto Insurance**
Ad auto.insure.com/NewYork ▼
Save on New York Auto Insurance.
Cheap Rates as Low as \$18.99/Month!
- \$29 Cheap Auto Insurance**
Ad nyc.usautoinsurance.com/ ▼
Cheapest NYC Car Insurance.
Lowest Rates From \$28.99 / Month!

At the bottom of the left column, there is a map snippet showing a location in Harlem, New York, with markers for "Nationwide Insurance - Steward Ins Agency Inc" and "Apollo Theater".

Chances to click can depends on the search query : **Contextual bandits**

Chances to click can depend on other ads: **Bandits with assortments (MNL-bandit)**

Examples: Dynamic Pricing

- A Seller with goods to price
- N possible discrete prices
- Observes sales (or no sale) only for the offered price
- Explore different prices or pick the best performing price so far?

Other considerations:

Continuous space of prices: [Continuum armed bandits](#)

Often involves inventory constraints: [Bandits with global constraints](#)

Reinforcement learning

- Limited feedback model with uncertainty generated by Markovian stochastic process
- Reward at time t is determined by the “action or arm” and “state” of the system
- At time t
 - Observe the current “state” of the system s_t
 - Take action a_t
 - Observe reward r_t , from fixed **unknown reward distribution**
 - System transitions to next state s_{t+1} with **unknown probability** $P(s_t, a_t, s_{t+1})$
- Maximize total reward or discounted reward

Reinforcement learning

- Trial-and-error
 - Explore-exploit
- Explore different actions, and observe reward and state transitions to learn
 - which actions have high reward in a given state
 - Which actions take you to good states
- Adapt exploration to past observations – learn from past mistakes
 - Limit exploration of bad states and bad actions

Game playing

- A computer algorithm playing game like Atari Breakout
- Maximize the score
- Need to make moves sequentially
- State: what you see on screen
- Limited feedback: Can observe only the outcome of move made
- Solve it by reinforcement learning – use only feedback, trial and error

