

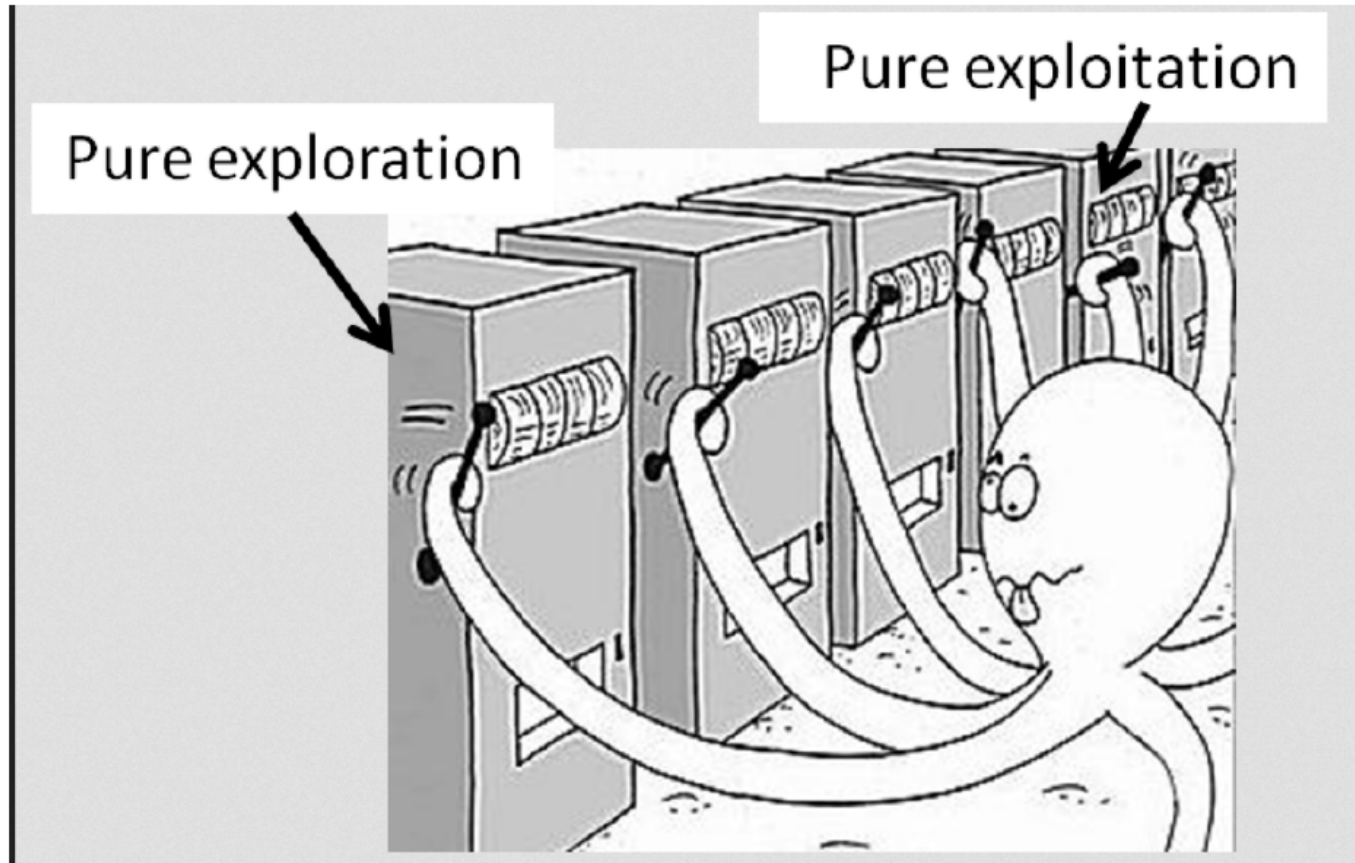
# LEC019 MAB I

VG441 SS2020

Cong Shi  
Industrial & Operations Engineering  
University of Michigan

# Basic RL for combining learning and decisions

- Multi-Armed Bandit Problem



# MAB

- Different machine generates different random rewards
- Gambler decides which slot machine to play with each token
- Maximize reward (\$\$)



# Online decision-making: learning while doing

- Online decision-making involves a fundamental choice:
  - Exploration: Gather more information
  - Exploitation: Make the best decision given current information



- The best long-term strategy may involve short-term sacrifices

# Example: Insufficient Exploration

1	2	3	4	5	6	7	8	
---	---	---	---	---	---	---	---	--



# Example: Insufficient Exploration

1	2	3	4	5	6	7	8	
	\$0		\$0	\$0				
\$5		\$5			\$5	\$5	\$5	...



# Example: Insufficient Exploration

1	2	3	4	5	6	7	8	
---	---	---	---	---	---	---	---	--

	\$0		\$0	\$0				
--	-----	--	-----	-----	--	--	--	--

\$5	\$5	\$5	\$5	\$5	\$5	\$5	\$5	...
-----	-----	-----	-----	-----	-----	-----	-----	-----



It turns out

always pays \$5/round

# Example: Insufficient Exploration

1	2	3	4	5	6	7	8	
---	---	---	---	---	---	---	---	--

	\$0		\$0	\$0				
--	-----	--	-----	-----	--	--	--	--

\$5	\$5	\$5	\$5	\$5	\$5	\$5	\$5	...
-----	-----	-----	-----	-----	-----	-----	-----	-----



It turns out



always pays **\$5**/round



pays **\$100** a quarter of the time  
(**\$25**/round on average)



# Example: Insufficient Exploration

1	2	3	4	5	6	7	8	
\$100	\$0	\$0	\$0	\$0	\$100	\$0	\$100	
\$5	\$5	\$5	\$5	\$5	\$5	\$5	\$5	...



It turns out



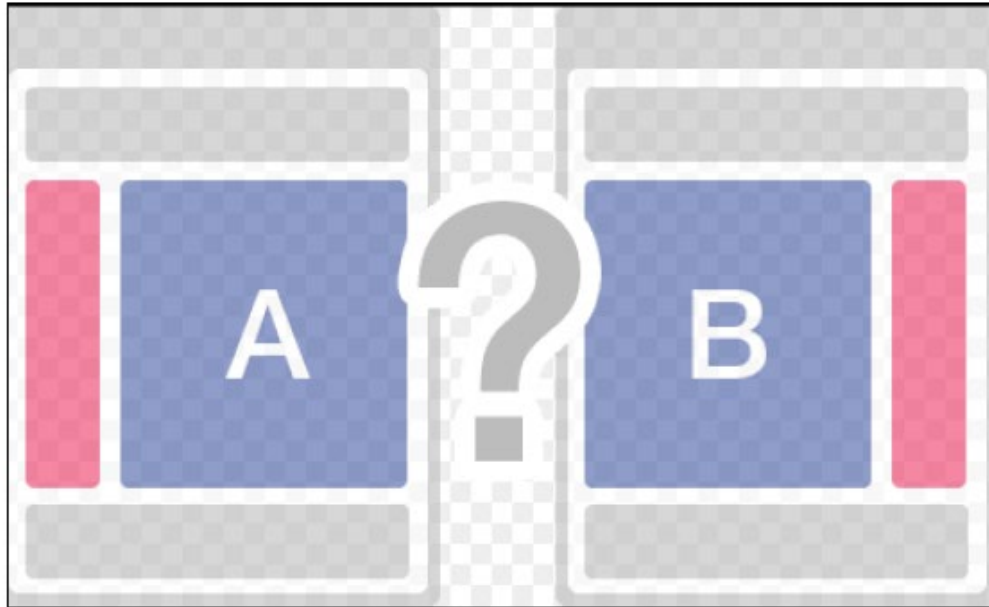
always pays **\$5**/round



pays **\$100** a quarter of the time  
(**\$25**/round on average)

# A/B Testing

- Exploration: Gather more information about which design is better
- Exploration: Show the best design to the customer



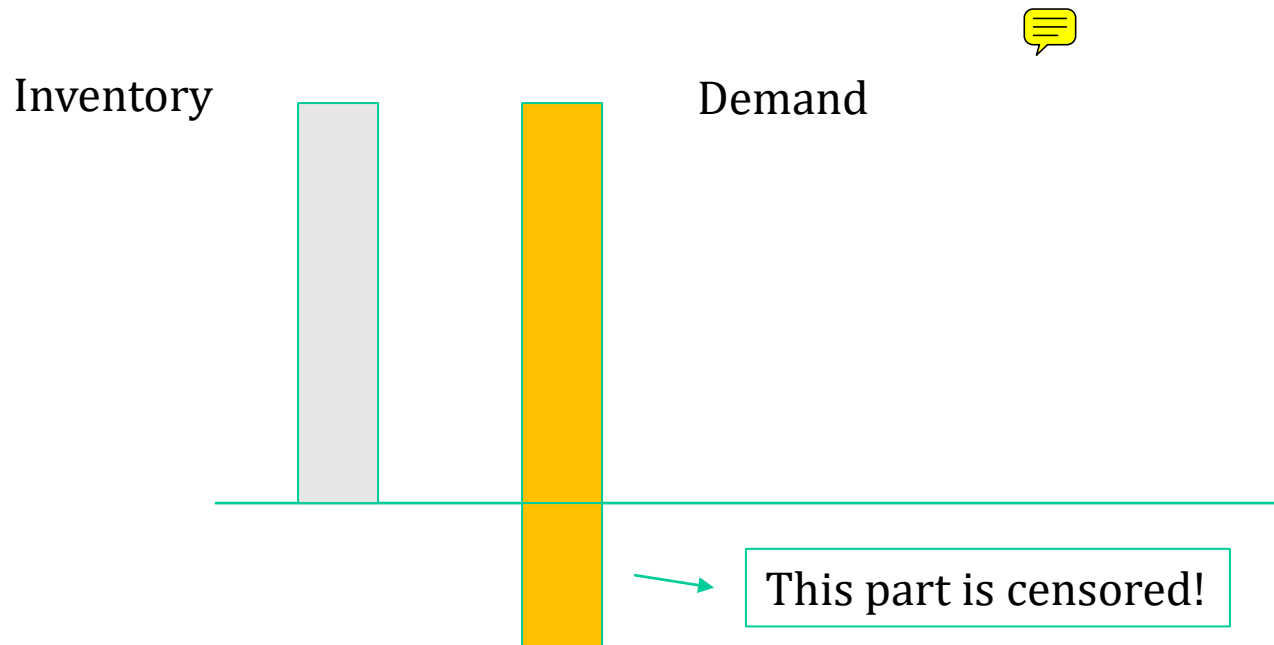
# Revenue Management

- Retailers are interested in finding an optimal (pricing) policy to max revenue
- Unknown relationship between price and customer's purchasing decision (demand distribution)
  - Exploration: Gather more information about customers behavior using different prices
  - Exploitation: Make the best price based on the current information



# Inventory Management

- Retailers are interested in finding an optimal (ordering) policy to min cost
- Unknown demand distribution (can only observe sales – censored demand)
  - Exploration: Order more to find out about true demand distribution
  - Exploitation: Order just right to minimize the cost



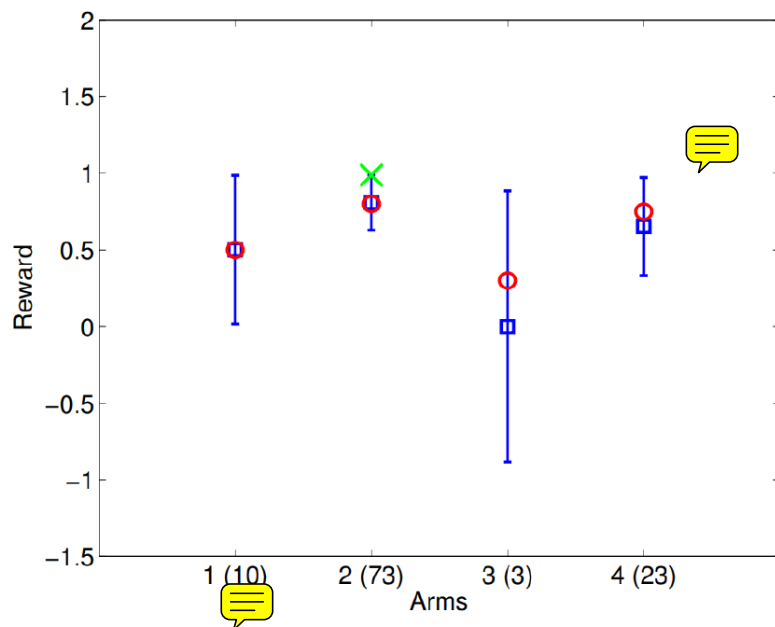
# Other Applications

- Clinical trials
- Recommender systems
- Advertising: what ad to put on a web-page?
- Auctions
- Financial portfolio design
- Crowdsourcing

# Many algorithms for MAB

## • $\epsilon$ -greedy algorithm

- Upper confidence bound (UCB)
  - Add confidence bonus to the estimated mean
  - If the estimator is reliable, add less; if not, add more



$$i_t = \arg \max_i \left[ \hat{\mu}_i + \underbrace{\sqrt{\frac{c \log t}{n_i}}}_{\text{ucb}_i} \right]$$

$$P\{|\bar{p} - p| \leq \delta\} \geq 1 - 2e^{-2n\delta^2}$$

当  $\delta$  取值为  $\sqrt{2 \ln T / n}$  时 (其中  $T$  表示有  $T$  个客人,  $n$  表示菜被吃过的次数), 可以得到

$$P\{|\bar{p} - p| \leq \sqrt{2 \ln T / n}\} \geq 1 - \frac{2}{T^4}$$

- Thompson sampling
  - Bayesian setup with a prior distribution over reward parameters
  - Choose the auction that maximizes the expected reward under posterior

# Online Network RM using TS

- ~\$300B industry with ~10% annual growth over the last 5 years
  - ♦ IBISWorldUS Industry Report; excludes online sales of traditionally brick & mortar stores

amazon.com®





priceline.com®

RueLaLa®

- Online retailers have additional information as compared to brick & mortar retailers, e.g. real-time customer purchase decisions (buy / no buy)
  - ♦ How can we use this information to develop a more effective revenue management strategy?

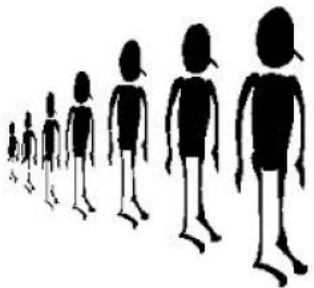
# Setting

- Finite selling horizon of  $T$  periods
  - One customer arrives per period
  - Sequentially observe customer purchase decisions
- Finite set of prices;  $i$ -th price denoted by  $p_i$  
- Unknown mean demand per price (“purchase probability”)  $d_i$
-  Given unlimited inventory and known demand, select price with highest revenue  $= p_i \times d_i$
- Challenges: unknown demand
- Exploration vs. Exploitation Tradeoff

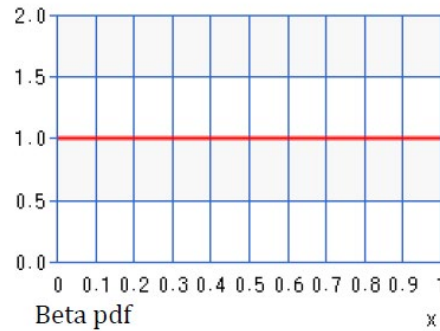
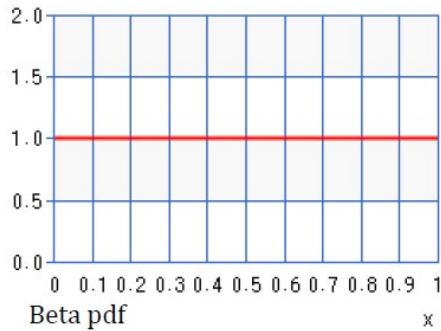


# RM-MAB


- Retailer decides...
  - Which price to offer to a customer
  - How many times to offer each price
  - In what order to offer prices to customers
- Learns demand at each price to max revenue



# RM-MAB



1. Customer arrives
2. Retailer samples  $\theta_1$  and  $\theta_2$  from current distributional estimation of  $d_1$  and  $d_2$
3. Retailer offers price that maximizes  $p_i \theta_i$
4. Customer makes purchase decision (according to  $d_i$ )
5. Retailer observes purchase decision and updates demand estimation

  $\hat{d}_1 \sim \text{Beta}(1,1)$        $\hat{d}_2 \sim \text{Beta}(1,1)$   
True (unknown)  $d_1 = 0.6$       True (unknown)  $d_2 = 0.3$

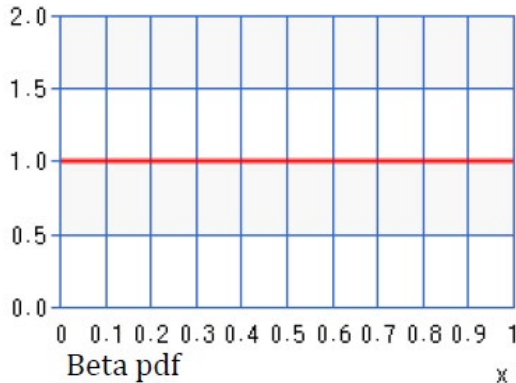


# RM-MAB

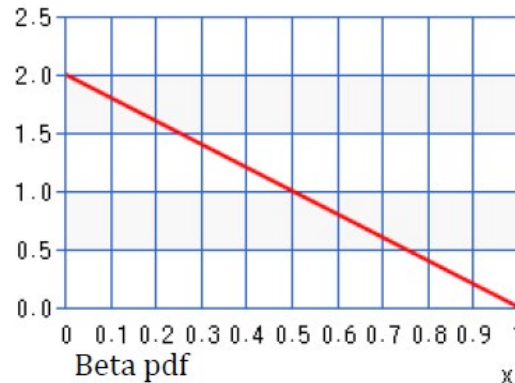


$$\theta_1 = 0.41, \theta_2 = 0.83$$

$$p_2\theta_2 > p_1\theta_1$$



$\hat{d}_1 \sim \text{Beta}(1,1)$   
True (unknown)  $d_1 = 0.6$



$\hat{d}_2 \sim \text{Beta}(1,1)$   
True (unknown)  $d_2 = 0.3$



update

$\hat{d}_2 \sim \text{Beta}(1, \mathbf{1} + \mathbf{1})$   
True (unknown)  $d_2 = 0.3$

Customer does not buy item 2

# RM-MAB

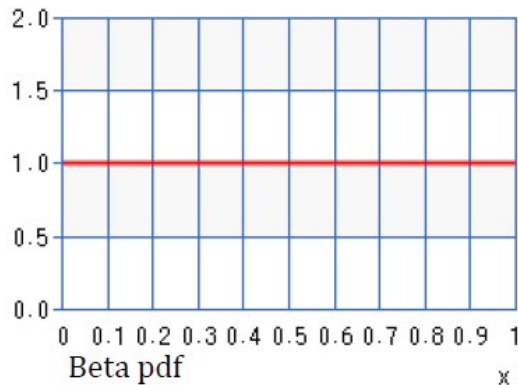


$$\theta_1 = 0.93, \theta_2 = 0.12$$

$$p_1\theta_1 > p_2\theta_2$$



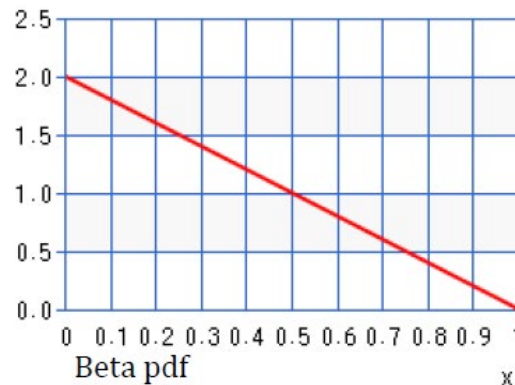
Customer buys item 1



$\hat{d}_1 \sim \text{Beta}(1,1)$   
True (unknown)  $d_1 = 0.6$



$\hat{d}_1 \sim \text{Beta}(1+1,1)$   
True (unknown)  $d_1 = 0.6$



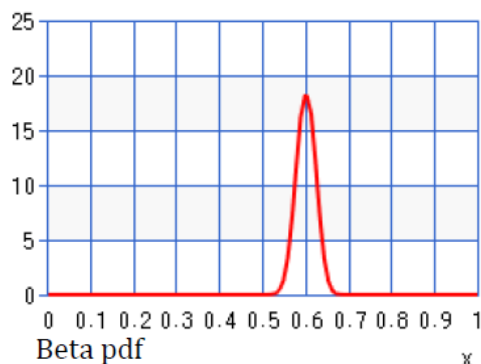
$\hat{d}_2 \sim \text{Beta}(1,2)$   
True (unknown)  $d_2 = 0.3$

update

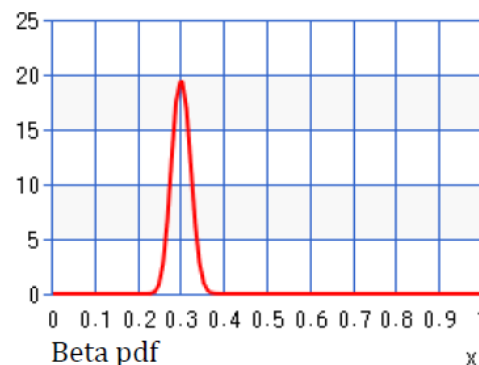
# RM-MAB: 2 Price Example

As each price is offered more times...

- Beta pdf converges to reflect true mean demand
- Will choose optimal price with high probability




$\hat{d}_1 \sim \text{Beta}(1 + \# \text{ "buy"}, 1 + \# \text{ "no buy"})$   
True (unknown)  $d_1 = 0.6$



$\hat{d}_2 \sim \text{Beta}(1 + \# \text{ "buy"}, 1 + \# \text{ "no buy"})$   
True (unknown)  $d_2 = 0.3$



# Advantages of Thompson Sampling

- Empirical and theoretical results show it's a highly competitive algorithm for unlimited inventory
- Easy to implement and understand
- Non-parametric 
- Continuous exploration & exploitation

**How do we incorporate inventory constraints?**

Key Tradeoffs:

- Exploration vs. Exploitation
- Explore at the cost of running out of inventory 




# RM-with inventory constraint

1. Customer arrives
2. Retailer samples  $\theta_1$  and  $\theta_2$
3. Retailer solves a deterministic LP to identify the optimal fraction of remaining customers to offer  $p_1$  and  $p_2$ , using
  - $\theta_1$  and  $\theta_2$
  - Remaining unsold inventory & customers
4. Retailer offers price  $p_i$  with probability based on fraction found in Step 3
5. Customer makes purchase decision
6. Retailer observes decision and updates  $\hat{d}_i$

# RM-with inventory constraint

$x_i =$  fraction of remaining customers  $(T-t)$  to offer price  $p_i$



  $\max_{x_1, x_2} \sum_{T-t} p_1 \theta_1 x_1 + p_2 \theta_2 x_2$

maximize revenue over remaining customers

$$\text{s.t. } x_1 + x_2 \leq 1$$

fraction of remaining customers  $\leq 1$

$$(T - t)(\theta_1 x_1 + \theta_2 x_2) \leq \text{Inv}(t)$$

expected inventory sold is upper-bounded by remaining inventory

$$x_1, x_2 \geq 0$$





# RM-with inventory constraint



$$\begin{aligned}\text{Regret} &= \mathbb{E}[\text{Revenue of Optimal Policy with Known Demand}] - \\ &\quad \mathbb{E}[\text{Revenue of Algorithm}] \\ &\leq \text{Upper Bound on Optimal Policy} - \\ &\quad \mathbb{E}[\text{Revenue of Algorithm}]\end{aligned}$$

---

## Theorem

Suppose the LP of the underlying true demand (i.e. benchmark) is nondegenerate. Then, for the modified Thompson Sampling with Inventory Algorithm,

$$\text{Regret}(T) \leq O(\sqrt{T} \log T \log \log T) = \tilde{O}(\sqrt{T})$$