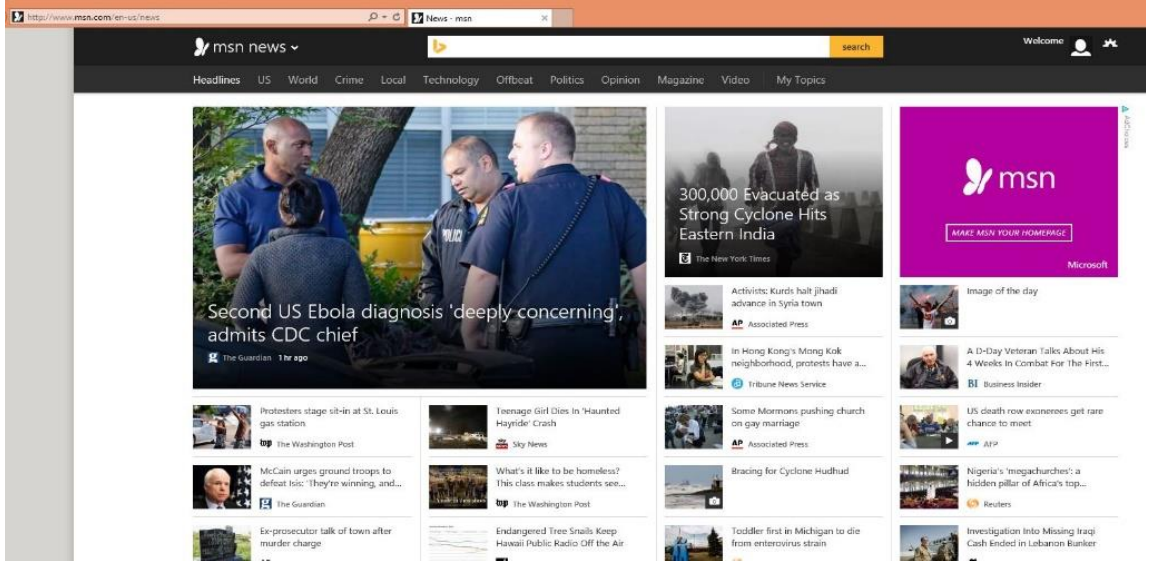
Contextual Bandits Overview

Alekh Agarwal
Microsoft Research NYC

Personalized news?



Repeatedly:

1. Observe features of user+articles
2. Choose a news article.
3. Observe click-or-not

Goal: Maximize fraction of clicks

Health advice?

Repeatedly:

1. Observe features of user+advice
2. Choose an advice.
3. Observe steps walked

Goal: Healthy behaviors (e.g. step count)

Other Real-world Applications

News Rec: [LCLS ‘10]

Ad Choice: [BPQCCPRSS ‘12]

Ad Format: [TRSA ‘13]

Education: [MLLBP ‘14]

Music Rec: [WWHW ‘14]

Robotics: [PG ‘16]

Wellness/Health: [ZKZ ’09, SLLSPM ’11, NSTWCSM ’14, PGCRRH ’14, NHS ’15, KHSBATM ‘15, HFKMTY ’16]

Contextual Bandits (CB)

For $t = 1,2,\dots,T$:

1. Observe features $x_{\mathrm{t}} \sim D_t$
2. Choose action $a_{\mathrm{t}} \in A$
3. Observe reward $r_{\mathrm{t}} \sim D_t(\cdot|x_t)$

Goal: Maximize net reward

$$\boldsymbol{E_{D_t}\left[\sum_{t=1}^{T} r_t\right]}$$

- $|A| = K, \; r_t \in [0,1]$

Adversarial and i.i.d.

i.i.d.

For $t = 1,2,\ldots,T$:

1. Observe features $x_{\mathrm{t}} \sim D$
2. Choose action $a_{\mathrm{t}} \in A$
3. Observe reward $r_{\mathrm{t}} \sim D(\cdot|x_t)$

Goal: Maximize net reward

$$\boldsymbol{E\sum_{t=1}^{T} r_t}$$

Adversarial

For $t = 1,2,\ldots,T$:

1. Observe features x_{t}
2. Simultaneously adversary picks $r_t \in [0,1]^K$
3. Choose action $a_{\mathrm{t}} \in A$
4. Observe reward $r_{\mathrm{t}}(a_t)$

Goal: Maximize net reward

$$\boldsymbol{E_{D_t}\left[\sum_{t=1}^{T} r_t(a_t)\right]}$$

How much reward is good?

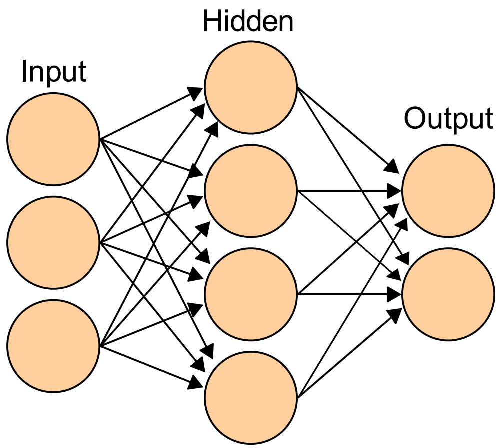
- Need a benchmark for comparison to our cumulative rewards

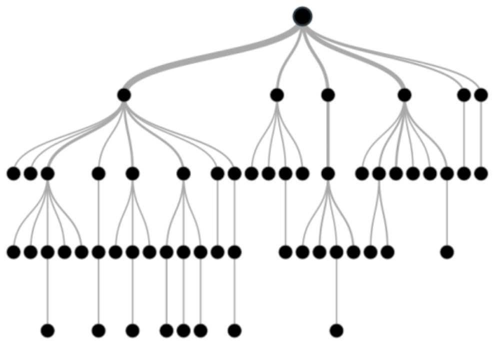
- **MAB:** Compare with the best fixed action in hindsight
  - **Tacit assumption:** A fixed action gets high rewards across all contexts
  - *e.g. same treatment to each patient, irrespective of their symptoms!*

- **EXP4:** Comparison with best expert
  - Good benchmark if we have a good expert

Policies

Policy maps features to actions.





Policy = Classifier that ***acts*.**

- chosen action = prediction of a classifier on the context

Use policies to pick actions in CB

How much reward is good?

- **CB:** Compare with the best fixed policy in a policy class
    - **Tacit assumption:** There is a policy which attains high reward in the class

- Pick an expressive class of policies to capture complex behaviors
- Allows taking different good actions on different contexts

- Limiting to a class restricts complexity for learning, like a hypothesis/concept class in supervised learning

Regret

$$Regret_T = \sum_{t=1}^{T} r_t - \sum_{t}^{T} \max_{\pi \in \Pi} r_t(\pi(x_t))$$

Best policy in hindsight

Connection to other learning settings

- MAB: Different benchmark makes CB harder and more useful
- Supervised learning: Wait for next lecture
- Reinforcement learning: Actions do not have long-term consequences on future contexts and rewards in CB.

Contextual Bandits(ish) Applications

News: Lihong Li, Wei Chu, John Langford, Robert E. Schapire: A contextual-bandit approach to personalized news article recommendation. WWW 2010.

Robotics: Lerrel Pinto, Abhinav Gupta: Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours. ICRA 2016: 3406-3413.

Music: Xinxi Wang, Yi Wang, David Hsu, Ye Wang: Exploration in Interactive Personalized Music Recommendation: A Reinforcement Learning Approach. TOMCCAP 11(1): 7:1-7:22 (2014).

Education: Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, Zoran Popovic, Offline policy evaluation across representations with applications to educational games. AAMAS 2014: 1077-1084.

Ad Format: Liang Tang, Rómer Rosales, Ajit Singh, Deepak Agarwal: Automatic ad format selection via contextual bandits. CIKM 2013: 1587-1594.

Ad Choice: Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, Ed Snelson: Counterfactual reasoning and learning systems: the example of computational advertising. JMLR 14(1): 3207-3260 (2013).

Wellness Contextual Bandits Work

P. Paredes, R. Gilad-Bachrach, M. Czerwinski, A. Roseway, K. Rowan and J. Hernandez, "Pop Therapy: Coping with stress through pop-culture," in Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, 2014.

I. Hochberg, G. Feraru, M. Kozdoba, S. Mannor, M. Tennenholtz, E. Yom-Tov (2016) "Encouraging Physical Activity in Diabetes Patients Through Automatic Personalized Feedback Via Reinforcement Learning Improves Glycemic Control" Diabetes Care 39(4): e59-e60

S. M. Shortreed, E. Laber, D. Z. Lizotte, S. T. Stroup, J. Pineau and S. A. Murphy, "Informing sequential clinical decision-making through reinforcement learning: an empirical study," Machine learning, vol. 84, no. 1-2, pp. 109-136, 2011.

I. Nahum-Shani, E. B. Hekler and D. Spruijt-Metz, "Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework," Health Psychology, vol. 34, p. 1209, 2015.

I. Nahum-Shani, S. S. Smith, A. Tewari, K. Witkiewitz, L. M. Collins, B. Spring and S. Murphy, "Just in time adaptive interventions (jitais): An organizing framework for ongoing health behavior support," Methodology Center technical report, 2014.

P. Klasnja, E. B. Hekler, S. Shiffman, A. Boruvka, D. Almirall, A. Tewari and S. A. Murphy, "Microrandomized trials: An experimental design for developing just-in-time adaptive interventions," Health Psychology, vol. 34, p. 1220, 2015.

Y. Zhao, M. R. Kosorok and D. Zeng, "Reinforcement learning design for cancer clinical trials," Statistics in medicine, vol. 28, no. 26, p. 3294, 2009.