Linear Bandits: Rich decision sets

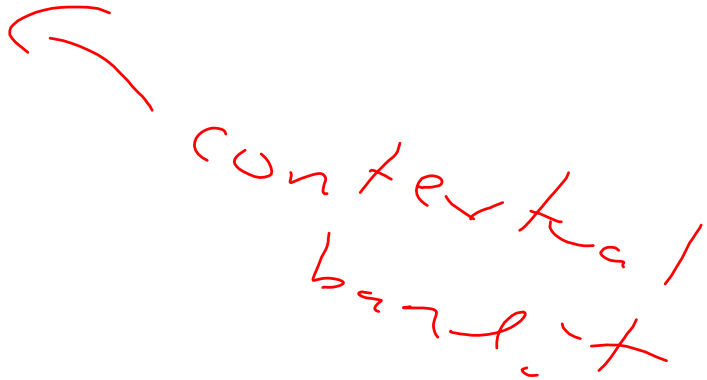
Sham M. Kakade

Machine Learning for Big Data
CSE547/STAT548

University of Washington

Bandits in practice: two major issues

- The decision space is very large.
  - Drug cocktails
  - Ad design
- We often have “side information” when making a decision
  - history of a user



More real motivations...

Clinical trials:





$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

- choose a treatment A_t for patient t
- observe a response $X_t \in \{0, 1\} : \mathbb{P}(X_t = 1) = \mu_{A_t}$
- Goal: maximize the number of patient healed

Recommendation tasks:

ν_1 ν_2 ν_3 ν_4 ν_5

- recommend a movie A_t for visitor t
- observe a rating $X_t \sim \nu_{A_t}$ (e.g. $X_t \in \{1, \ldots, 5\}$)

Linear bandits

- An additive effects model.
- Suppose each round we take a decision $x \in \mathcal{D} \subset \mathcal{R}^d$.
  - x is paths on a graph.
  - x is a feature vector of properties of an ad
  - x is a which drugs are being taken
- Upon taking action x, we get reward r, with expectation:

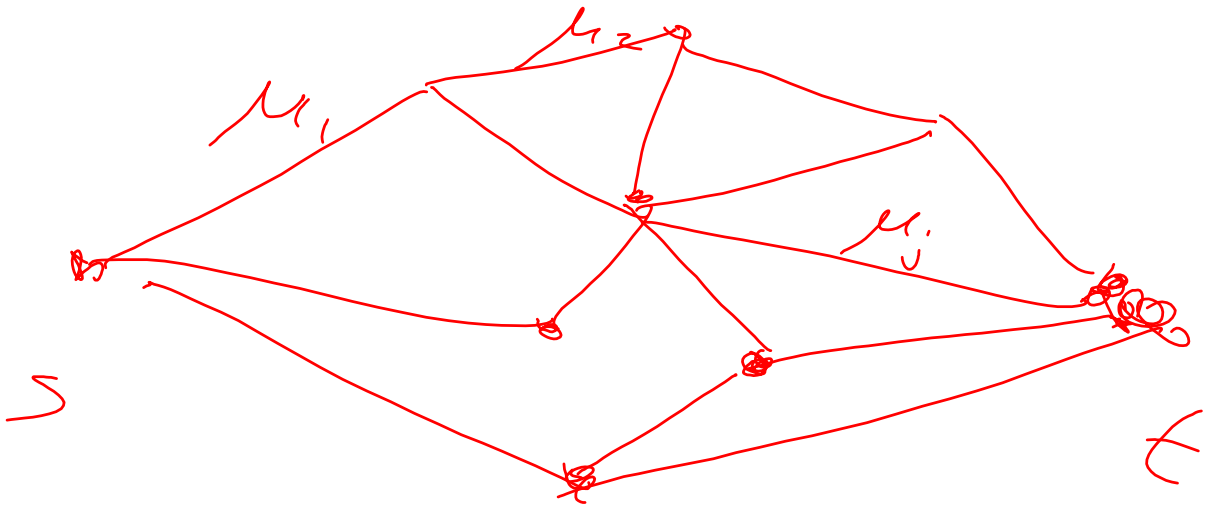
$$\mathbb{E}[r|x] = \mu^\top x$$

- only d unknown paramters (and "effectively" 2^d actions)
- W desire an algorithm $\mathcal{A}$ (mapping histories to deicsions), which has low regret.

$$\mu^\top x_* - \mathbb{E}[\mu^\top x_t|\mathcal{A}] \leq ??$$

(where x_* is the best decision)

Example: Shortest paths...



$x \in \{0,1\}^E$

x is a path in G

Algorithm Idea

- again, let's think of optimism in the face of uncertainty
- we observed some $r_1, \ldots r_{t-1}$, and have taken $x_1, \ldots x_{t-1}$.
- Questions:
  - what is an estimate of the reward of $\mathbb{E}[r|x]$ and what is our uncertainty?
  - what is an estimate of μ and what is our uncertainty?

Regression!

- Define:

$$A_t := \sum_{\tau<t} x_\tau x_\tau^\top + \lambda I, \;\; b_t := \sum_{\tau<t} x_\tau r_\tau$$

- Our estimate of μ

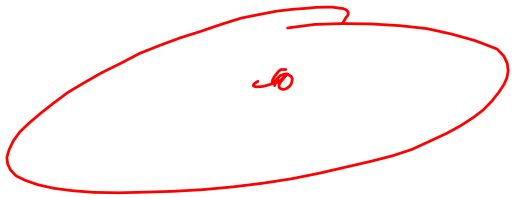
$$\hat{\mu}_t = A_t^{-1} b_t$$

- Confidence of our estimate:

$$\|\mu - \hat{\mu}_t\|_{A_t}^2 \leq \mathcal{O}(d \log t)$$

$$(\mu - \hat{\mu}_t)^\top A_t \; (\mu_t - \hat{\mu}_t)$$

LinUCB



- Again, optimism in the face of uncertainty.
- Define:

$$B_t := \{\nu | \|\nu - \hat{\mu}\|_{A_t}^2 \leq \mathcal{O} d \log t\}$$

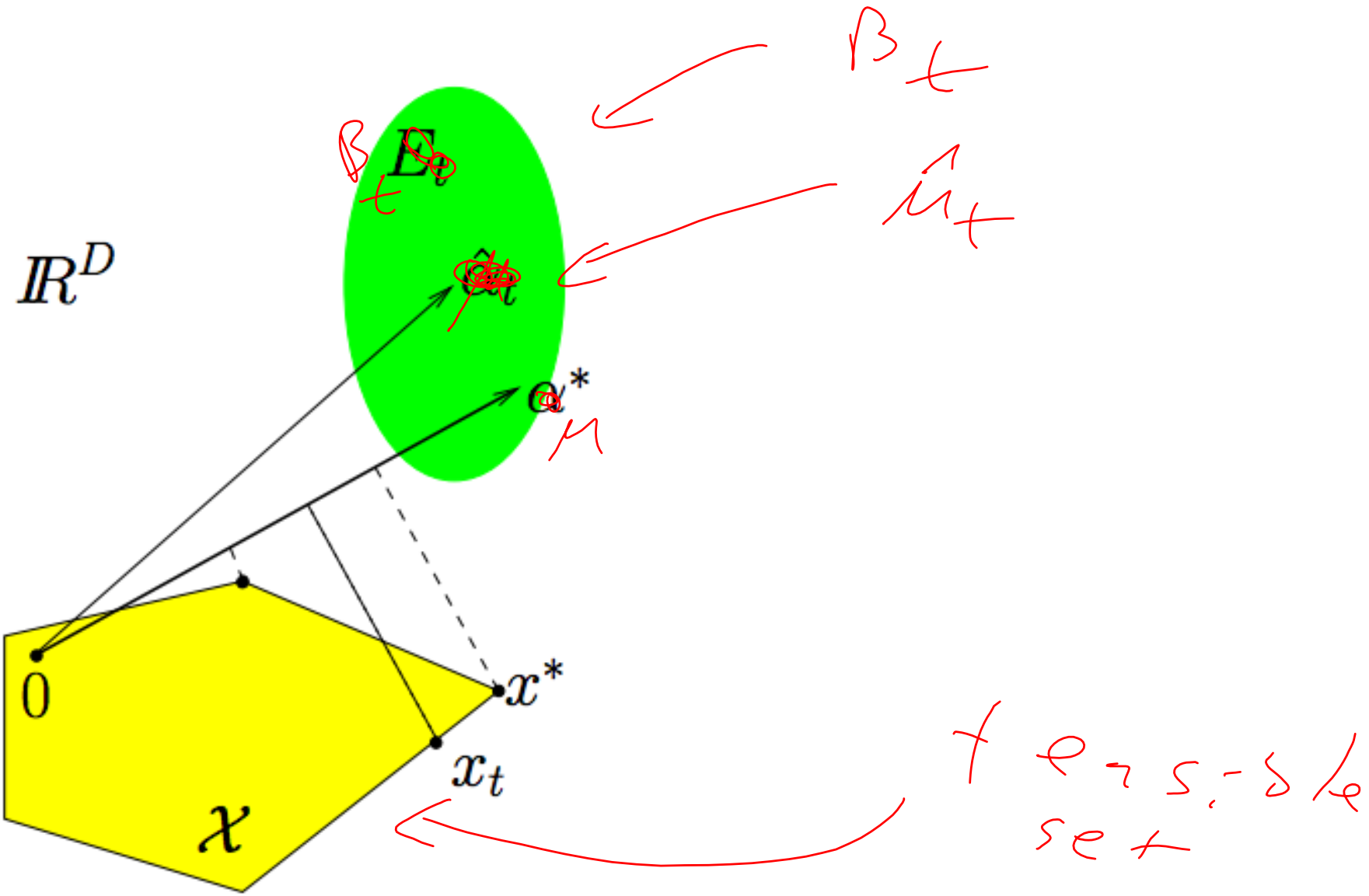
- (Lin UCB) take action:

$$x_t = \text{argmax}_{x \in \mathcal{D}} \max_{\nu} \nu^\top x$$

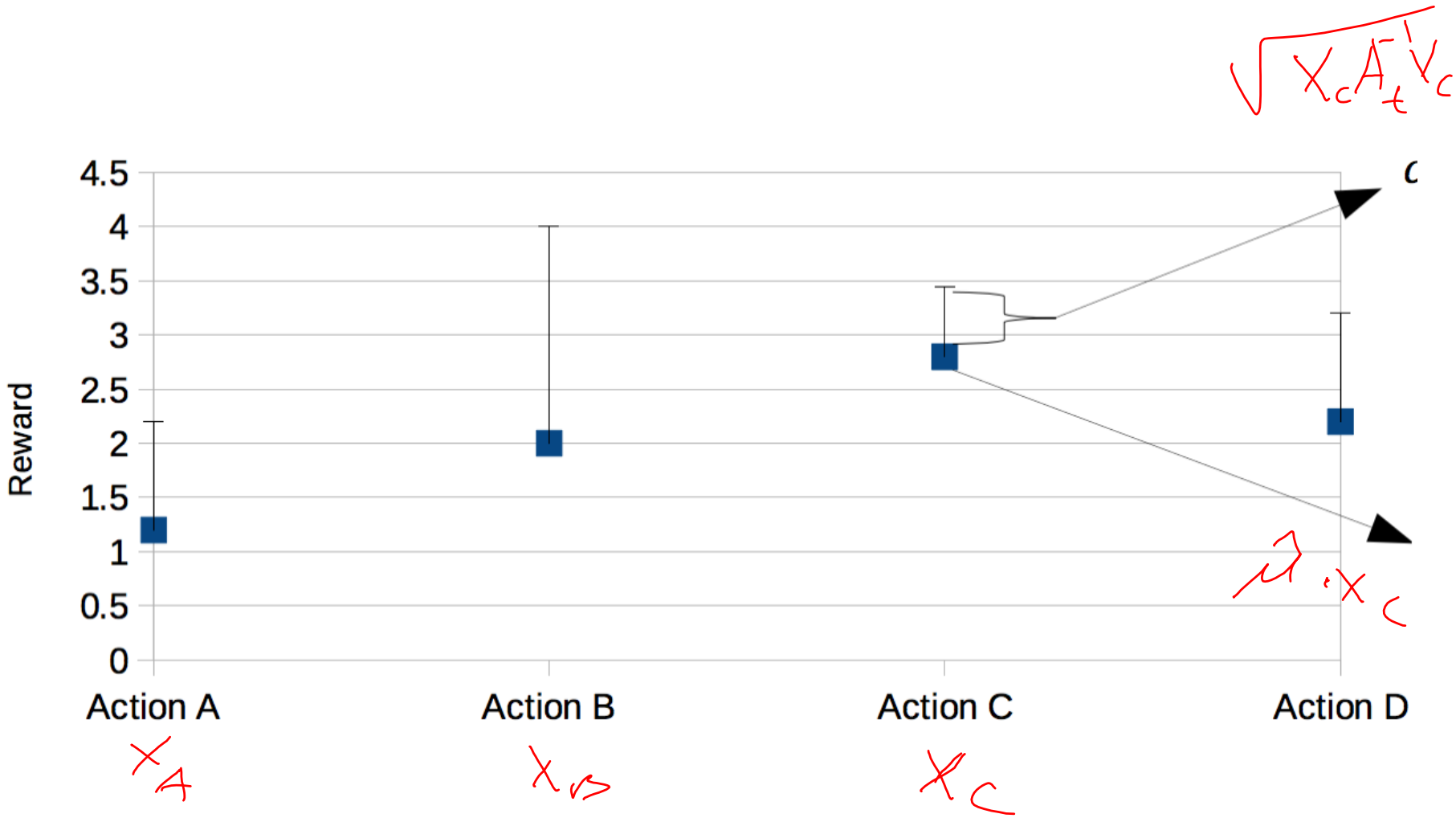
  then update A_t, B_t, b_t, and $\hat{\mu}_t$.
- Equivalently, take action:

$$x_t = \text{argmax}_{x \in \mathcal{D}} \ \hat{\mu}^\top x + (d \log t)\sqrt{x A_t^{-1} x}$$

LinUCB: Geometry



LinUCB: Confidence intervals



LinUCB

- Regret bound:

$$T\mu^\top x_* - \sum_t \mathbb{E}[\mu^\top x_t | \mathcal{A}] \leq {}^*(d\sqrt{T})$$

  (this is the best possible, up to log factors).
- Compare to $O(\sqrt{KT})$
  - Independent of number of actions.
  - k-arm case is a special case.
- Thompson sampling: This is a good algorithm in practice.

Proof Idea...

- Stats: need to show that B_t is a valid confidence region.
- Geometric lemma: The regret is upper bounded by the:

$$\log \frac{\text{volume of posterior cov}}{\text{volume of prior cov}}$$

- Then just find the worst case log volume change.

Dealing with context...

Dealing with context...

Acknowledgements

- http://gdrro.lip6.fr/sites/default/files/JourneeCOSdec2015-Kaufman.pdf
- https://sites.google.com/site/banditstutorial/
- http://www.yisongyue.com/courses/cs159/lectures/LinUCB.pdf