Online Learning
Other topics

Yevgeny Seldin

Evaluation of Bandit Algorithms

Evaluation of bandit algorithms in practice

- Challenge: previously unobserved actions or (state,action) pairs

- Deployment
  - Risky and time-consuming

- Environment simulation
  - Requires a good simulator
    - This may be very hard or even impossible to produce
    - If we have a good simulator, we probably already have a solution to the problem

Evaluation of bandit algorithms in practice

- Offline evaluation for i.i.d. problems
  1. Use full information data where possible and relevant
  2. "Importance-weighting" of logged limited feedback data
     - Requires randomized sampling in the logging policy with non-zero probability for taking all the (potentially relevant) actions
     - Requires logging the sampling distribution (to do importance-weighting)
     - Variance of the estimates scales with $\frac{1}{p_{\text{logging}}(a)}$
- Evaluation in the adversarial regime
  - Generally impossible
  - Sparring

Alternative algorithms for bandits

Alternative algorithms for i.i.d. bandits

- UCB-style algorithms
  - kl-UCB (based on kl inequality)
  - UCB-V (based on Empirical Bernstein or Unexpected Bernstein inequality)
- Thompson sampling (Bayesian approach)
- Subsampling
- Best-of-both-worlds algorithms

Variations of EXP3 – high probability regret bound

- EXP3
  - $p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$
  - $\tilde{\ell}_{t,a} = \frac{\ell_{t,a}\mathbb{I}(A_t=a)}{p_t(a)}$
  - $\mathbb{E}[R_T] = O\left(\sqrt{KT \ln K}\right)$

- EXP3-IX: high-probability regret guarantee
  - $\tilde{\ell}_{t,a} = \frac{\ell_{t,a}\mathbb{I}(A_t=a)}{p_t(a)+\frac{\eta_t}{2}}$
  - $\mathbb{P}\left(R_T \geq O\left(\sqrt{KT \ln K} \ln \frac{1}{\delta}\right)\right) \leq \delta$

Variations of EXP3 – best-of-both-worlds

- EXP3
  - $p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$
  - $\tilde{\ell}_{t,a} = \frac{\ell_{t,a}\mathbb{I}(A_t=a)}{p_t(a)}$
  - $\mathbb{E}[R_T] = O\left(\sqrt{KT\ln K}\right)$

- EXP3++: best-of-both-worlds
  - $\tilde{p}_t(a) = \left(1 - \sum_a \varepsilon_t(a)\right)p_t(a) + \varepsilon_t(a)$
  - $\varepsilon_t(a) = \theta\left(\frac{\ln t}{t\,\widehat{\Delta}_t(a)^2}\right)$, where $\widehat{\Delta}_t(a)$ is a lower confidence bound on the gap
  - $\mathbb{E}[R_T] = O\left(\sqrt{KT\ln K}\right)$
  - $\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{(\ln T)^2}{\Delta(a)}\right)$

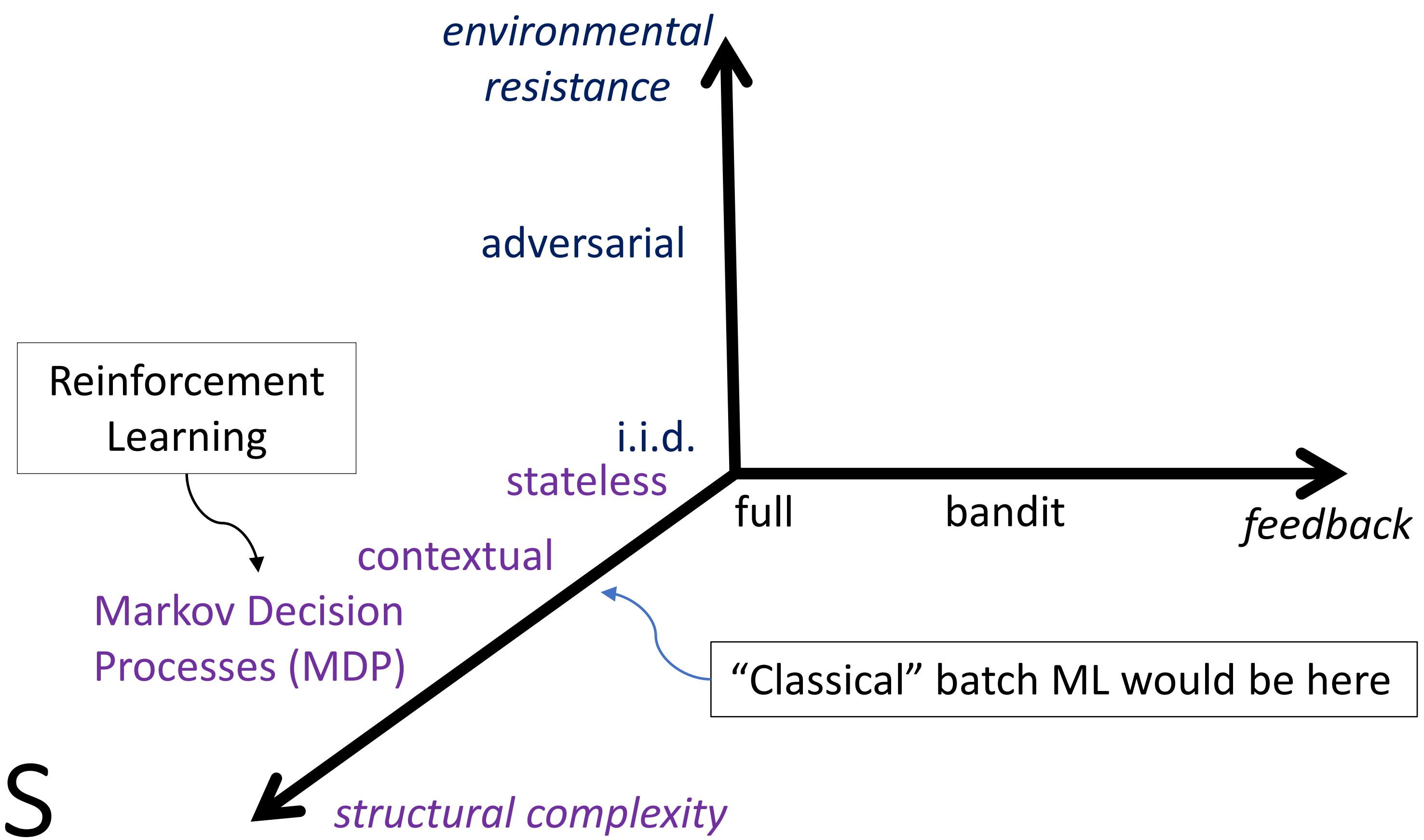
Adversarial bandits: alternative regularization

- EXP3
  - $p_t = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}} = \arg\min_p \langle p, L_{t-1} \rangle \underbrace{+ \frac{1}{\eta_t} \sum_a p_a \ln p_a}_{\substack{\text{Regularization} \\ \text{Negative entropy}}}$

- Tsallis-INF – the ultimate algorithm: Best-of-both-worlds and minimax optimal
  - $p_t = \arg\min_p \langle p, L_{t-1} \rangle \underbrace{- \frac{1}{\eta_t} \sum_a \sqrt{p_a}}_{\substack{\text{Regularization} \\ \text{Tsallis entropy}}}$

  - Adversarial: $\mathbb{E}[R_T] = O\left(\sqrt{KT}\right)$
  - I.I.D.: $\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{\ln T}{\Delta(a)}\right)$

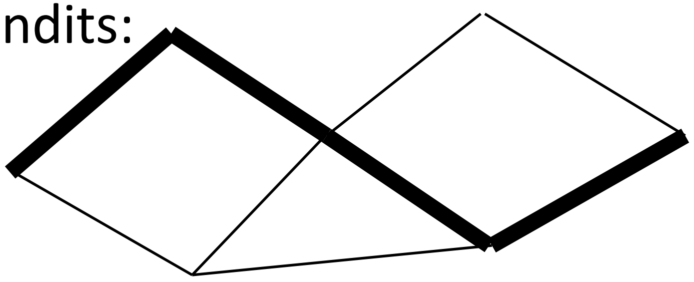
Other Settings



Structure forms: (Generalized) Linear Bandits

Linear Bandits:

- $r_t = \langle \bar{A}_t, \bar{\theta}_* \rangle + \xi_t$
- $\bar{A}_t \in \mathcal{D}_t$
- Special cases:
  - $\mathcal{D} = \{(1,0,\dots,0),\dots,(0,\dots,0,1)\}$ - multiarmed bandits
  - $\mathcal{D}_t = \{\phi(c_t, a) \colon a \in \{1,\dots,K\}\}$ - contextual bandits
  - Combinatorial (semi-)bandits:
  - Cascading bandits



Generalized Linear Bandits:

- $r_t = f\left(\langle \bar{A}_t, \bar{\theta}_* \rangle\right) + \xi_t$

Feedback forms

- From full to limited: paid observations, decoupled exploration, graph feedback, …
- Dueling Bandits
  - Relative comparison of pairs arms, but not their true value
    - Would you like fish or chicken?
  - Very useful for implicit information collection from user feedback
- Ranking
  - Selection from a ranked list
- Partial Monitoring
  - Separation between observations and losses
  - Example: dynamic pricing

Environment forms

- Contaminated stochastic
- Stochastically constrained adversarial

Bandit variations

- Bandits with switching costs
- Recharging/recovering bandits
- Rotting bandits
- Bandits with knapsacks
- ….

Delayed feedback

Alternative objectives

- We have studied regret minimization
  - Cumulative loss of actions along the way

- Pure Exploration / Best arm identification / Experiment design
  - Find the best action as fast as possible
  - Losses along the way are not counted

Summary

- An infinite world of exciting problem formulations