Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems

Konstantina Christakopoulou

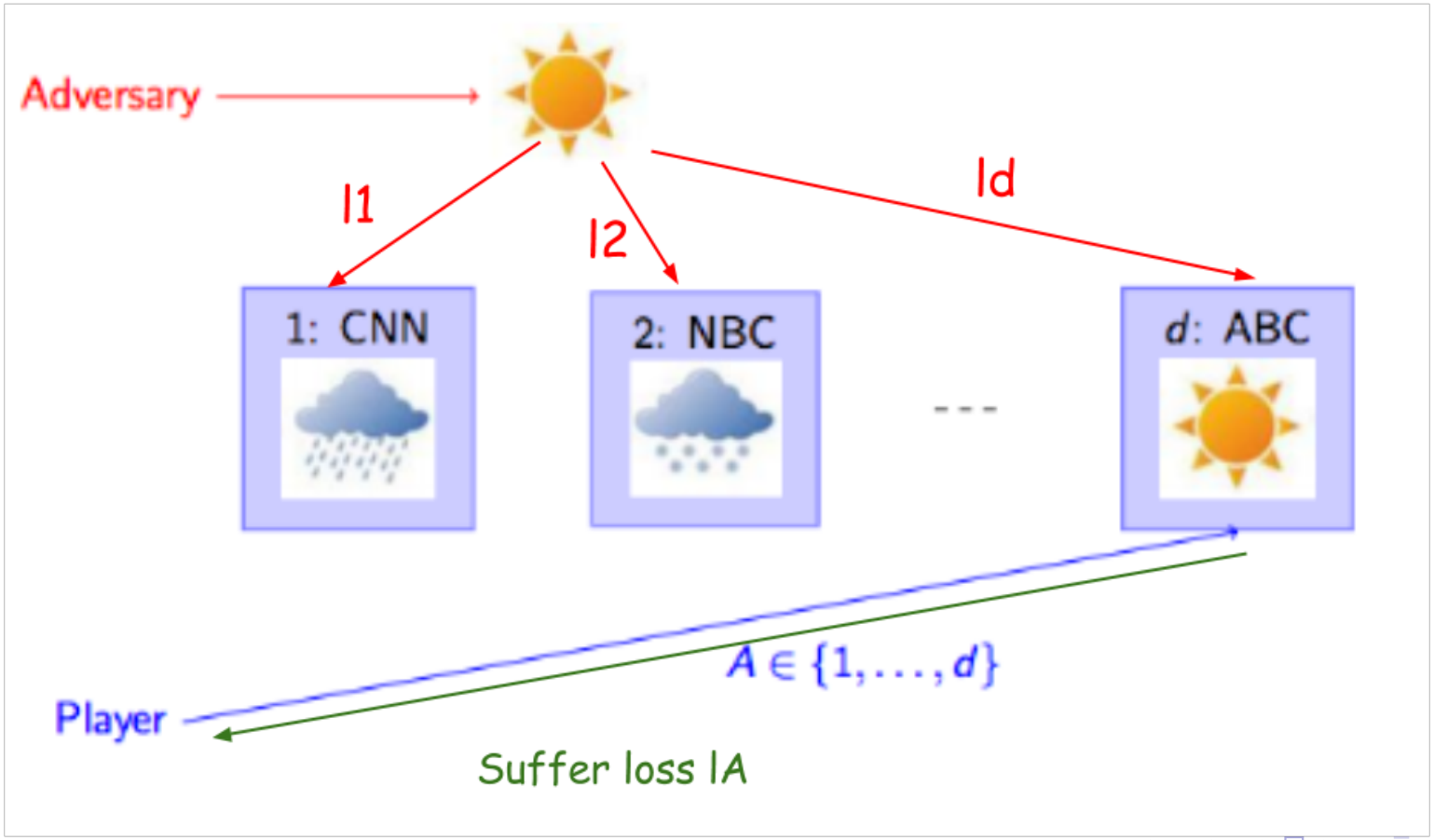
University of Minnesota

christa@cs.umn.edu

March 12, 2014

Online Learning with Full vs. Bandit information

- Full information setting: Observe $\ell_1, \ell_2, \ldots, \ell_d$
- Bandit setting: Observe ℓ_A



The Stochastic Bandit Setting

K arms, K probability distributions $\nu_1, \nu_2, \ldots, \nu_K$ on $[0, 1]$
At each round $t = 1, \ldots, (n)$:

- The learner chooses $I_t \in 1, \ldots, K$
- Given I_t, the world reacts with reward $X_{I_t,t} \sim \nu_{I_t}$ independently from the past.

Goal: Minimize regret

$$R_n = \max_{i=1,\ldots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}$$

Pseudo-regret:

$$\overline{R_n} = \max_{i=1,\ldots,K} \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right]$$

Exploration - Exploitation !

Optimism in the face of uncertainty

- $\Delta_i = \mu^* - \mu_i$ and $T_i(s) = \sum_{t=1}^{s} \mathbb{1}_{I_t=i}$
  $\overline{R_n} = \left(\sum_{i=1}^{K} \mathbb{E}\, T_i(n)\right) \mu^* - \mathbb{E} \sum_{i=1}^{K} T_i(n)\mu_i = \sum_{i=1}^{K} \Delta_i \mathbb{E}\, T_i(n)$

1. When uncertain $\rightarrow$ consider best possible world and choose best arm
2. UCB: Construct upper bound estimate for μ_i at fixed confidence level and choose $I_t = \arg\max_i \left\{ \hat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{\alpha \ln(t)}{2T_i(t-1)}} \right\}$
   - α big $\rightarrow\uparrow$ exploration VS α small $\rightarrow\uparrow$ exploitation
   - $\overline{R_n} \leq O(\frac{1}{\Delta}\alpha \ln(n))$
3. Lower bound for any strategy where $\mathbb{E}\, T_i(n) = o(n^\alpha)$ for any set of Bernoulli reward distributions: $\lim_{n\to\infty} \inf \frac{\overline{R}_n}{\ln n} \geq \sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathrm{kl}(\mu_i, \mu_i^*)}$
4. other exploration techniques: $\epsilon-$ greedy, epoch greedy, Thompson sampling, etc.

The Adversarial Bandit Setting

K arms

At each round $t = 1, \ldots, (n)$:

- The learner chooses $I_t \in 1, \ldots, K$
- At the same time, the adversary selects gain vector $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0, 1]^K$
- The learner receives the reward $g_{I_t,t}$, while the gains of the other arms are not received.

Exp3: Exponential weights for Exploration and Exploitation

Let $(\eta_t)_{t\in\mathbb{N}}$ non increasing sequence of real numbers and p_1 uniform distribution over $\{1,\dots,K\}$
At each round $t = 1,\dots,n$:

- Draw arm I_t from p_t
- For each arm $i = 1,\dots,K$ compute estimated loss $\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i}$ and update estimated cumulative loss $\tilde{L}_{i,t} = \tilde{L}_{i,t-1} + \tilde{\ell}_{i,s}$
- Compute new probability distribution over arms p_{t+1} where:

$$p_{i,t+1} = \frac{\exp(-\eta_t \tilde{L}_{i,t})}{\sum_{k=1}^{K} \exp(-\eta_t \tilde{L}_{k,t})}$$

Unbiased estimator Exponential reweighting trick

Regret Bounds on Exp3

Exp3 satsifies $\overline{R_n} \leq \frac{K}{2} \sum_{t=1}^{n} \eta_t + \frac{\ln K}{\eta_n} \rightarrow$ tradeoff!

Theorem (Pseudo-regret of Exp3)

If $\eta = \sqrt{\frac{2 \ln K}{nK}}$ *then* $\overline{R_n} \leq \sqrt{2nK \ln K}$.

If Exp3 is run with $\eta_t = \sqrt{\frac{\ln K}{tK}}$, *then* $\overline{R_n} \leq 2\sqrt{nK \ln K}$.

Comparison with online learning: $O(\sqrt{n \log K})$

Exp3.P

1. Trick no.1: to derive high probability bounds Bias in gain estimate
2. Trick no.2: Add some extra uniform exploration

$\eta \in \mathbb{R}^+$ and $\gamma, \beta \in [0,1]$. p_1 uniform distribution over $\{1, \ldots, K\}$
At each round $t = 1, \ldots, n$:

- Draw arm I_t from p_t
- For each arm $i = 1, \ldots, K$ compute estimated gain $\tilde{g}_{i,t} = \frac{g_{i,t} \mathbb{1}_{I_t = i} + \beta}{p_{i,t}}$ and update estimated cumulative gain $\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$
- Compute new probability distribution over arms p_{t+1} where:

$$p_{i,t+1} = (1-\gamma) \frac{\exp(\eta \tilde{G}_{i,t})}{\sum_{k=1}^{K} \exp(\eta \tilde{G}_{k,t})} + \frac{\gamma}{K}$$

High Probability Bounds on Exp3.P and Lower Bound

Theorem (High probability bound for Exp3.P)

For any $\delta \in (0,1)$, *if Exp3.P is run with* $\beta = \sqrt{\frac{\ln K\delta^{-1}}{nK}}, \eta = 0.95\sqrt{\frac{\ln K}{nK}}$, $\gamma = 1.05\sqrt{\frac{K \ln K}{n}}$, *with probability at least* $1-\delta$, $R_n \leq 5.15\sqrt{nK \ln(K\delta^{-1})}$

For any confidence level, if $\beta = \sqrt{\frac{\ln K}{nK}}$ *then with probability at least* $1-\delta$,

$R_n \leq \sqrt{\frac{nK}{\ln K}} \ln \delta^{-1} + \sqrt{nK \ln(K)}$

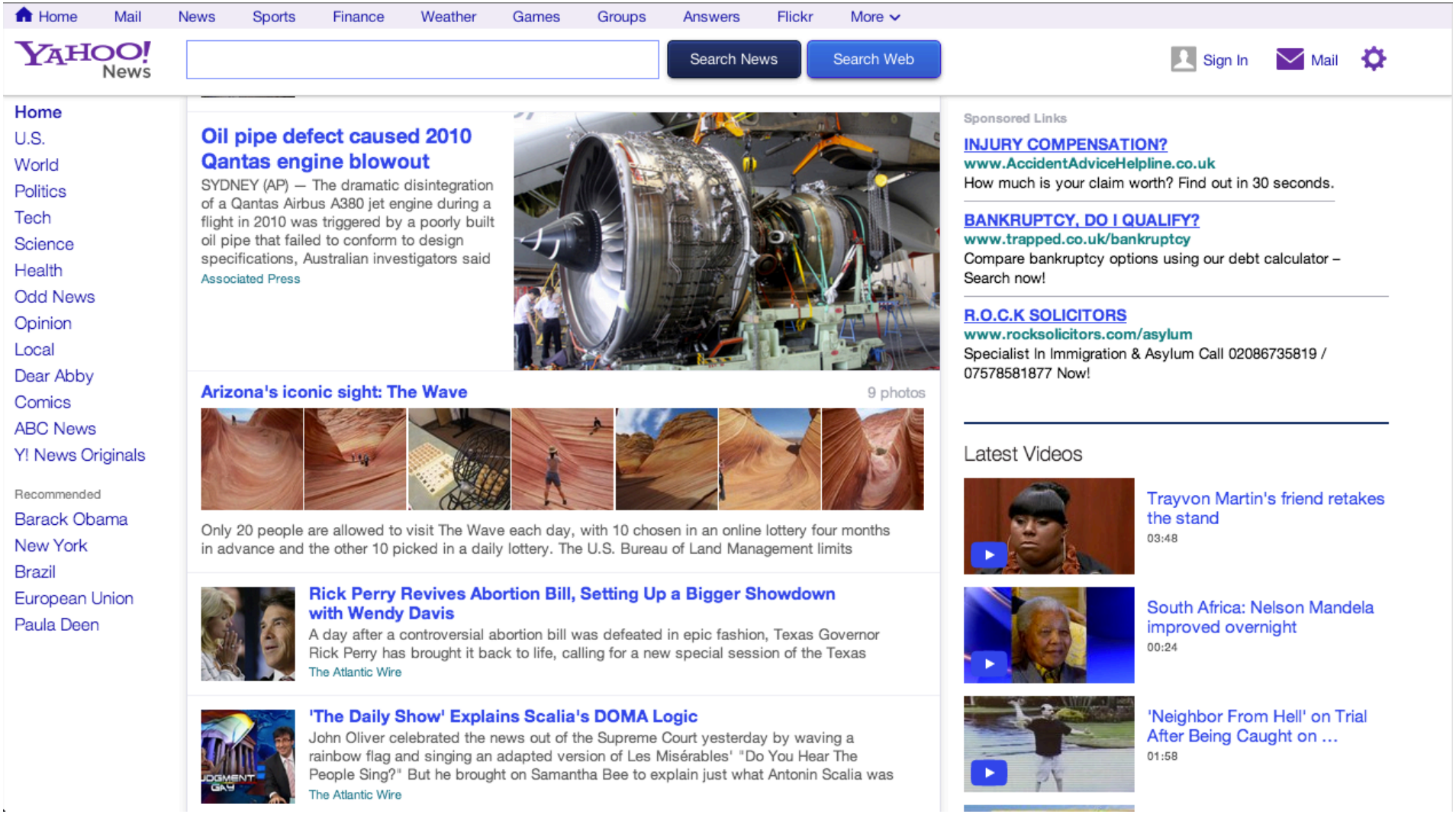
Bound $\mathbb{E}R_n$

Theorem (Minimax Lower Bound)

$$\inf_{\text{forecasters}} \sup_{\text{adversaries}} \left(\max_{i=1,\ldots,K} \mathbb{E} \sum_{t=1}^{n} Y_{i,t} - \mathbb{E} \sum_{t=1}^{n} Y_{I_t,t} \right) \geq \frac{1}{20}\sqrt{nK}$$

$\mathbb{E}R_n \geq \overline{R_n}$ Results unimprovable up to a logarithmic bound

The Contextual Bandit Setting



The Contextual Bandit Setting

At each round $t = 1, \ldots, (n)$:

- The world produces some context $s \in S$
- The learner chooses $I_t \in 1, \ldots, K$
- The world reacts with reward $X_{I_t,t}$

Goal: Learn a good policy for choosing actions given context.

$$\overline{R_n}^S = \max_{g:S\to\{1,\ldots,K\}} \mathbb{E}\left[\sum_{t=1}^{n} \ell_{I_t,t} - \sum_{t=1}^{n} \ell_{g(s_t),t}\right]$$

1. Idea: Run a separate instance of Exp3 on each distinct context $\rightarrow$ S-Exp3 $\rightarrow \overline{R_n}^S \leq \sqrt{2n|S|K \ln K}$
2. The Expert case: Run Exp3 on N randomized policies (experts) $\rightarrow O(\sqrt{nN \log N})$. So, Exp4: $O(\sqrt{nK \log N})$

Contextual bandit: The expert case

Exp4 Algorithm

N experts. Let $(\eta_t)_{t\in\mathbb{N}}$ non increasing sequence of real numbers and q_1 uniform distribution over $\{1,\dots,N\}$
At each round $t = 1,\dots,n$:

- Get expert advice $\xi_t^1,\dots,\xi_t^N$, where ξ_t^j probability distribution over arms.
- Draw arm I_t from $p_t = (p_{1,t},\dots,p_{K,t})$ where $p_{i,t} = (1-\gamma)\mathbb{E}_{j\sim q_t}\xi_{i,t}^j + \frac{\gamma}{K}$
- For each arm $i = 1,\dots,K$ compute estimated loss $\tilde{\ell}_{i,t} = \frac{\ell_{i,t}\mathbb{1}_{I_t=i}}{p_{i,t}}$
- Compute estimated loss for each expert $\tilde{y}_{j,t} = \mathbb{E}_{i\sim\xi_t^j}\tilde{\ell}_{i,t}$
- Update estimated cumulative loss for each expert $j = 1,\dots,N$: $\tilde{Y}_{j,t} = \sum_{s=1}^t \tilde{y}_{j,s}$
- Compute new probability distribution over experts q_{t+1}: $q_{j,t+1} = \frac{\exp(-\eta_t \tilde{Y}_{j,t})}{\sum_{k=1}^N \exp(-\eta_t \tilde{Y}_{k,t})}$

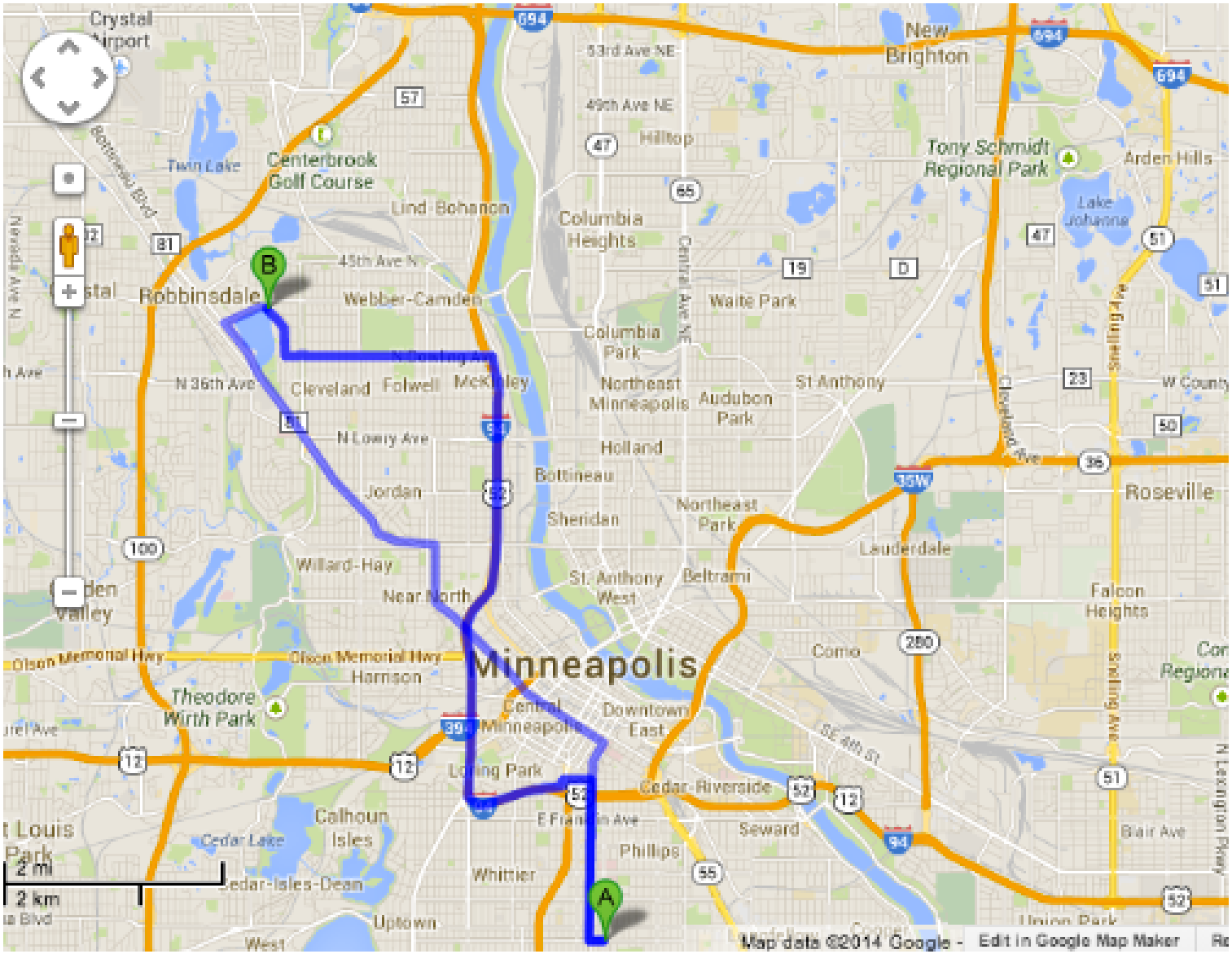
Competing against the best context set: Run Exp4 with mixing coefficient γ using experts $\equiv$ instances of S-Exp3. Each instance run on different $S_\theta, \theta\in\Theta \to$ $\overline{R_n}^{\Theta} = \max_{\theta\in\Theta}\max_{g:S_\theta\to\{1,\dots,K\}}\mathbb{E}\left[\sum_{t=1}^n \ell_{I_t,t} - \sum_{t=1}^n \ell_{g(s_{\theta,t}),t}\right]$ $O(\sqrt{\ln|\Theta|})$.

Stochastic Contextual Bandit: Policy evaluation

- $F = \{f : S \to \{1, \ldots, K\}\}$ set of policies
- Supervised learning: observed: (x_t, ℓ_t) vs bandit: observed data $(x_t, \ell_{I_t,t})$
- $f^* = \arg\inf_{f \in F} \mathbb{E}_{(x,I):\ \text{iid} \sim D} \ell_{f(x)} \to$ bound $\sum_{t=1}^{n} \ell_{I_t,t} - n\ell_D(f^*)$
- - For $f : X \to \{1, 2\}$, $K = 2$ arms and $VC(F) = d$:
    1. For n' rounds, choose arms uniformly at random
    2. Build $F' \subseteq F$ s.t for any $f \in F$ there is exactly one $f' \in F'$ $f(x_t) = f'(x_t)$, $t = 1, \ldots, n'$
    3. For $t = n' + 1, \ldots, n$ play Exp4 using policies of F' as experts.

  VE: per round regret is $\sqrt{\frac{d}{n}} \equiv$ (optimal rate for supervised VC).
- different views on contextual bandits: e.g Banditron ($O(n^{2/3})$) for multi class case.

The Linear Bandit Setting



The Linear Bandit Setting

n rounds, action set: compact set $K \subset \mathbb{R}^d$, unknown $L \subset \mathbb{R}^d$

At each round $t = 1, \ldots, (n)$:

- The learner chooses $x_t \in K$
- The adversary chooses ℓ_t from some fixed and unknown subset L of $\mathbb{R}^d$
- The learner incurs and observes the loss $x_t^T \ell_t$

Goal: Minimize the cumulative pseudo regret:

$$\overline{R_n} = \mathbb{E} \sum_{t=1}^{n} x_t^T \ell_t - \min_{x \in K} \sum_{t=1}^{n} x^T \ell_t$$

Expanded Exponential weights strategy (Exp2)

Assumptions: (i) Bounded scalar loss $|x^T \ell| \leq 1$, (ii) Finite sets $K, |K| = N$

$$p_t(x) = (1-\gamma)\frac{\exp(-\eta \sum_{s=1}^{t-1} \langle x, \tilde{\ell}_t \rangle)}{\sum_{y \in K} \exp(-\eta \sum_{s=1}^{t-1} \langle y, \tilde{\ell}_t \rangle)} + \gamma \sum_{i=1}^{M} \mu_i \mathbb{1}_{x=u_i}$$

1. Idea: Importance sampling! If $x_t \sim p_t$, $\tilde{\ell}_t = \left(\mathbb{E}_{x \sim p_t}(xx^T)\right)^{-1} x_t x_t^T \ell_t$
   $\mathbb{E}_{x_t \sim p_t} \tilde{\ell}_t = \mathbb{E}_{x_t \sim p_t}(\mathbb{E}_{x \sim p_t}(xx^T)^{-1} x_t x_t^T \ell_t = \ell_t$ and $\mathbb{E}_{x_t \sim p_t} \tilde{\ell}_t^T \mathbb{E}_{x \sim p_t} = d$
2. Choose $P_t = \sum_{x \in K} p_t(x) x \otimes x \rightarrow \tilde{\ell}_t = P_t^{-1} \langle x_t, \ell_t \rangle x_t$
3. Choose exploration distribution μ

John's Theorem

$K \subset \mathbb{R}^d$ convex set. If the ellipsoid E of minimal volume enclosing K is the unit ball in some norm derived from a scalar product $\langle ., . \rangle$, then there exists $M \leq d(d+1)/2 + 1$ contact points $u_1, \ldots, u_M$ between E and K, and $\mu \in \Delta_M$ (the simplex of dimension $M-1$), such that $x = d \sum_{i=1}^{M} \mu_i \langle x, u_i \rangle u_i, \forall x \in \mathbb{R}^d$

$\overline{R_n} \leq 2\gamma n + \frac{\ln N}{\eta} + \eta n d$. For $\eta = \sqrt{\frac{\ln N}{3nd}}$ and $\gamma = \eta d$, $\overline{R_n} \leq 2\sqrt{3nd \ln N}$

Online Stochastic Mirror Descent (OSMD)

- Assume K is convex set $\rightarrow$ same regret for x_t and at random $\tilde{x}_t (\mathbb{E}[\tilde{x}_t] = x_t)$
- Parameters: F Legendre on $\overline{D} \supset conv(K)$ ((i) strictly convex and continuous first partial derivatives, (ii) $\lim_{x \to \overline{D} \setminus D} \|\nabla F(x)\| = \infty$)
- Bregman divergence $D_F(x, y) = F(x) - F(y) - (x - y)^T \nabla F(y)$
- sampling scheme $\pi : Conv(K) \rightarrow \Delta(K)$ s.t $\mathbb{E}_{X \sim \pi(a)} X = a$

OSMD

For each round $t = 1, 2, \ldots, n$:

1. Play random perturbation $\tilde{x}_t$ of x_t and observe $\ell_t(\tilde{x}_t)$
2. Compute random estimate $\tilde{g}_t$ of $\nabla \ell_t(x_t)$
3. $w_{t+1} = \nabla F^*(\nabla F(x_t) - \eta \tilde{g}_t)$
4. Project: $x_{t+1} = \arg\min_{y \in K} D_F(y, w_{t+1})$
5. $p_{t+1} = \pi(x_{t+1})$

Special cases of OSMD

1. For semi-bandit feedback: Let set of arms $C \subset \{0,1\}^d$, set of linear loss functions $L = [0,1]^d$, actions $u_t \in C, \|u\|_1 = m, m \leq d$
   - OSMD with Negative entropy ($F(x) = \sum_{i=1}^d x_i \ln x_i - \sum_{i=1}^d x_i$): OSMD on $K = \text{Conv}(C)$ with $\tilde{x}_t = u_t$ and $\tilde{\ell}_t(i) = \frac{\ell_t(i)u_t(i)}{x_t(i)}$. $\overline{R_n} \leq \sqrt{2mdn \ln \frac{d}{m}}$
   - OSMD with 0-potential (ω potential function $F_\psi(x) = \sum_{i=1}^d \int_\omega^{x_i} \psi^{-1}(s)ds$): Choose $\psi(x) = (-x)^{-q}$. For $q = 2, \overline{R_n} \leq 2\sqrt{2mdn}$ (optimal bound)

Special cases of OSMD

1. For semi-bandit feedback: Let set of arms $C \subset \{0,1\}^d$, set of linear loss functions $L = [0,1]^d$, actions $u_t \in C, \|u\|_1 = m, m \leq d$
   - OSMD with Negative entropy ($F(x) = \sum_{i=1}^d x_i \ln x_i - \sum_{i=1}^d x_i$): OSMD on $K = \text{Conv}(C)$ with $\tilde{x_t} = u_t$ and $\tilde{\ell_t}(i) = \frac{\ell_t(i) u_t(i)}{x_t(i)}$. $\overline{R_n} \leq \sqrt{2mdn \ln \frac{d}{m}}$
   - OSMD with 0-potential (ω potential function $F_\psi(x) = \sum_{i=1}^d \int_\omega^{x_i} \psi^{-1}(s)ds$): Choose $\psi(x) = (-x)^{-q}$. For $q = 2, \overline{R_n} \leq 2\sqrt{2mdn}$ (optimal bound)
2. For bandit feedback: Let compact set $K \subset \mathbb{R}^d$, losses $L \subseteq \mathrm{R}^d$, bounded scalar loss $|x^T \ell| \leq 1$
   - OSMD for euclidean ball ($K = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$): $\tilde{x_t} = x_t / \|x_t\|$ if $\xi_t = 1$ else $\epsilon_t e_{I_t}$ (ξ_t Bernoulli r.v, I_t uniformly at random, ϵ_t Rademacher r.v.) $\tilde{\ell_t} = d(1 - \xi_t) \frac{\tilde{x_t}^T \ell_t}{1 - \|x_t\|} \tilde{x_t} \rightarrow$ OSMD on $K' = (1-\gamma)K$ with $F(x) = -\ln(1 - \|x\|) - \|x\| \rightarrow R_n \leq 3\sqrt{dn \ln n}$ (optimal bound)

Minimax Regret bounds

	Bounded loss	**Combinatorial setting**
Full / Semi-bandit	$\sqrt{dn}$	$d\sqrt{n}$
Bandit	$d\sqrt{n}$	$d^{3/2}\sqrt{n}$

Table : Minimax regret bounds

- Obtained by Exp2 with John's exploration
- Obtained by OSMD with negative entropy
- Cannot get the optimal Best-known bound by Exp2: $O(d^2\sqrt{n})$

The Non - Linear Bandit setting

Loss: non-linear function of arms.

- Two-point feedback: $O(\sqrt{n})$ OSMD with $F = \frac{1}{2}\|\|^2$. Observe stochastic estimate $\tilde{g}_t(x_t)$ of $\nabla \ell_t(x_t)$:

$$\tilde{g}_t(x_t) = \frac{d}{2\delta}(\ell_t(X_t^+) - \ell_t(X_t^-))S$$

  where S r.v uniform in the unit sphere, $\delta > 0$ fixed, $X_t^+ = x_t + \delta S$, $X_t^- = x_t - \delta S$

- One point feedback: $O(n^{3/4})$

$$\tilde{g}_t(x) = \frac{d}{\delta}(\ell_t(x + \delta S)S$$

- 1-D stochastic case (Arms: points in $[0,1]$): $O(\sqrt{n}\log n)$
  $\mu : [0,1] \to [0,1]$ unimodal, not necessarily convex. Assume $|\mu(x) - \mu(x')| \geq C_L|x - x'|$ and $|\mu(x) - \mu(x')| \leq C_H|x - x'|$ (Lipschitz) $\to$ use stochastic golden search.

References

Sebastien Bubeck and Nicolo Cesa-Bianchi

Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems

Foundations and Trends in Machine Learning (2012)

Thank you!