Tuning bandit algorithms in stochastic environments

Jean-Yves Audibert, CERTIS - Ecole des Ponts
Remi Munos, INRIA Futurs Lille
Csaba Szepesvári, University of Alberta

The 18th International Conference on Algorithmic Learning Theory
October 3, 2007, Sendai International Center, Sendai, Japan

Contents

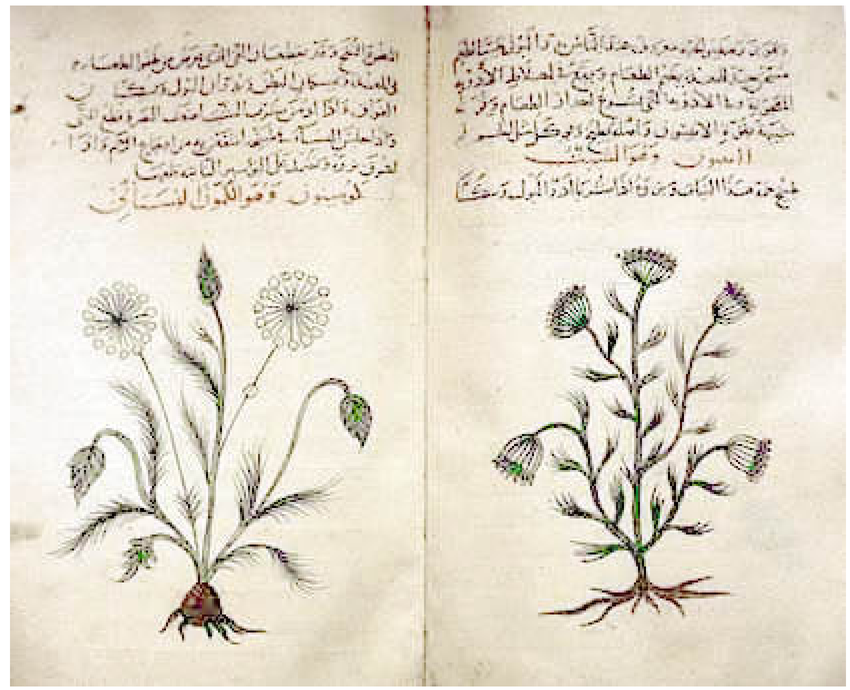
- Bandit problems
- UCB and Motivation

- Tuning UCB by using variance estimates
- Concentration of the regret
- Finite horizon – finite regret (PAC-UCB)

- Conclusions

Exploration vs. Exploitation

- Two treatments
- Unknown success probabilities
- Goal:
  - find the best treatment while losing the smallest number of patients
- Explore or exploit?



Playing Bandits

- Payoff is 0 or 1

- Arm 1:

  0 , 1 , 0 , 0 , X_{15}, X_{16}, X_{17}, …

- Arm 2:

  1 , 1 , 0 , 1 , 1 , 1 , X_{27}, …

Exploration vs. Exploitation: Some Applications

- Simple processes:
  - Clinical trials
  - Job shop scheduling (random jobs)
  - What ad to put on a web-page
- More complex processes (memory):
  - Optimizing production
  - Controlling an inventory
  - Optimal investment
  - Poker

Bandit Problems – "Optimism in the Face of Uncertainty"

- Introduced by Lai and Robbins (1985) (?)
- i.i.d. payoffs
  - $X_{11}, X_{12}, \ldots, X_{1t}, \ldots$
  - $X_{21}, X_{22}, \ldots, X_{2t}, \ldots$
- Principle:
  - Inflated value of an option = maximum expected reward that looks "quite" possible given the observations so far
  - Select the option with best inflated value

Some definitions

Now: $t=11$
$T_1(t-1) = 4$
$T_2(t-1) = 6$
$I_1 = 1$, $I_2 = 2$, …

- Payoff is 0 or 1
- Arm 1:

  0 , **1** , **0** , **0** , X_{15}, X_{16}, X_{17}, …

- Arm 2:

  1 , **1** , **0** , **1** , **1** , **1** , X_{27}, …

$$\hat{R}_n \stackrel{\text{def}}{=} \sum_{t=1}^{n} X_{k^*,t} - \sum_{t=1}^{n} X_{I_t,T_{I_t}(t)}$$

Parametric Bandits [Lai&Robbins]

- $X_{it} \sim p_{i,\theta_i}(\cdot)$, θ_i unknown, t=1,2,…
- Uncertainty set:
  "Reasonable values of θ given the experience so far"

  $U_{i,t}=\{\theta \mid p_{i,\theta}(X_{i,1:T_i(t)})$ is "large" mod $(t,T_i(t))\}$

- Inflated values:

  $Z_{i,t}=\max\{ E_\theta \mid \theta\in U_{i,t} \}$
- Rule:

  $I_t = \arg\max_i Z_{i,t}$

Bounds

- Upper bound:

$$\mathbb{E}\left[T_j(n)\right] \le \left(\frac{1}{D(p_j \| p^*)} + o(1)\right) \log(n)$$

- Lower bound:
  If an algorithm is uniformly good then..

$$\mathbb{E}\left[T_j(n)\right] \ge \left(\frac{1}{D(p_j \| p^*)} - o(1)\right) \log(n)$$

UCB1 Algorithm (Auer et al., 2002)

- Algorithm: UCB1(b)
  1. Try all options once
  2. Use option *k* with the highest index:

$$\hat{\mu}_{kt} + \sqrt{2b^2 \frac{\log(t)}{T_k(t-1)}}$$

- Regret bound:
  - R_n: Expected loss due to not selecting the best option at time step *n*. Then:

$$\mathbb{E}[R_n] \leq 8 \left(\sum_{k \in \mathrm{Bad}} \frac{b^2}{\Delta_k} \right) \log(n) + O(1)1$$

Problem #1

When $b^2 \gg \sigma^2$, regret should scale with σ^2 and not b^2!

UCB1-NORMAL

- Algorithm: UCB1-NORMAL
  1. Try all options once
  2. Use option k with the highest index:

$$\hat{\mu}_{kt} + \sqrt{16\hat{\sigma}_{kt}^2 \frac{\log(t)}{T_k(t-1)}}$$

- Regret bound:

$$\mathbb{E}\left[R_n\right] \leq 8 \left(\sum_{k \in \mathrm{Bad}} \frac{32\sigma_k^2}{\Delta_k} + \Delta_k \right) \log(n) + O(1)$$

Problem #1

- The regret of UCB1(b) scales with O(b^2)
- The regret of UCB1-NORMAL scales with O(σ^2)

  … but UCB1-NORMAL assumes normally distributed payoffs
- UCB-Tuned(b):

$$\hat{\mu}_{kt} + \sqrt{\min\left(\frac{b^2}{4}, \tilde{\sigma}^2_{kt}\right)\frac{\log(t)}{T_k(t-1)}}$$

  - Good experimental results
  - No theoretical guarantees

UCB-V

- Algorithm: UCB-V(b)
  1. Try all options once
  2. Use option k with the highest index:

$$\hat{\mu}_{kt} + \sqrt{2.4\tilde{\sigma}^2_{kt}\frac{\log(t)}{T_k(t-1)}} + \frac{3b\log(t)}{T_k(t-1)}$$

- Regret bound:

$$\mathbb{E}[R_n] \leq 10\left(\sum_{k\in\text{Bad}}\frac{\sigma_k^2}{\Delta_k} + 2b\right)\log(n)$$

Proof

- The "missing bound" (hunch.net):

$$|\hat{\mu}_t - \mu| \leq \sqrt{\frac{\tilde{\sigma}_t \log(3\delta^{-1})}{t}} + \frac{3b \log(3\delta^{-1})}{t}$$

- Bounding the sampling times of suboptimal arms (new bound)

Can we *decrease* exploration?

- Algorithm: UCB-V(b,ζ,c)
  1. Try all options once
  2. Use option k with the highest index:

$$\hat{\mu}_{kt} + \sqrt{2\zeta\tilde{\sigma}^2_{kt}\frac{\log(t)}{T_k(t-1)}} + c\frac{3b\log(t)}{T_k(t-1)}$$

- Theorem:
  - When $\zeta<1$, the regret will be polynomial for some bandit problems
  - When $c\zeta<1/6$, the regret will be polynomial for some bandit problems

Concentration bounds

- Averages concentrate:

$$\left|\frac{S_n}{n} - \mu\right| \leq O\left(\sqrt{\frac{\log(\delta^{-1})}{n}}\right)$$

- Does the regret of UCB* concentrate?

$$\left|\frac{R_n}{n} - \mu\right| \leq ??$$

RISK??

$$\left|\frac{R_n}{\mathbb{E}[R_n]} - 1\right| \leq ??$$

Logarithmic regret implies high risk

- **Theorem:**
  Consider the pseudo-regret
  $$R_n = \sum_{k=1}^{K} T_k(n)\, \Delta_k.$$
  Then for any $\zeta > 1$ and $z > \gamma \log(n)$,
  $$P(R_n > z) \leq C\, z^{-\zeta}$$
  (Gaussian tail: $P(R_n > z) \leq C \exp(-z^2)$)
- **Illustration:**
  - Two arms; $\Delta_2 = \mu_2 - \mu_1 > 0$.
  - Modes of law of R_n at $O(\log(n))$, $O(\Delta_2 n)$!

Only happens when the support of the second best arm's distribution overlaps with that of the optimal arm

Finite horizon: PAC-UCB

- **Algorithm:** PAC-UCB(*N*)
  1. Try all options ones
  2. Use option *k* with the highest index:

$$\hat{\mu}_{kt} + \sqrt{2\tilde{\sigma}^2_{kt}\frac{L_t}{T_k(t-1)}} + \frac{3bL_t}{T_k(t-1)},$$

$$L_t = \log(NK(T_k(t-1)+1))$$

- **Theorem:**
  - At time N with probability 1-1/N, suboptimal plays are bounded by O(log(K N)).
  - Good when N is known beforehand

Conclusions

- Taking into account the variance lessens dependence on the a priori bound b
- Low expected regret => high risk
- PAC-UCB:
  - Finite regret, known horizon, exponential concentration of the regret
- Optimal balance? Other algorithms?
- Greater generality: look up the paper!

Thank you!

Questions?

References

- Optimism in the face of uncertainty: Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6:4–22.
- UCB1 and more: Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.
- Audibert, J., Munos, R., and Szepesvári, Cs. (2007). Tuning bandit algorithms in stochastic environments, ALT-2007.