Conservative Bandits

Yifan Wu, Roshan Shariff, Tor Lattimore, Csaba Szepesvári
{ywu12, roshan.shariff, tlattimo, szepesva}@ualberta.ca
UNIVERSITY OF ALBERTA
DEPARTMENT OF COMPUTING SCIENCE

Introduction

We introduce:

- A new **multi-armed bandit problem:** challenge of exploring new strategies while maintaining fixed baseline of revenue.
- For **stochastic problem:** new algorithms satisfying minimum revenue constraint at every step; problem-dependent guarantees on their regret with respect to the optimal action.
- **Regret lower bounds** for stochastic problem, showing our algorithms are almost optimal.
- For **adversarial problem**: high-probability regret bounds showing penalty due to modifying existing algorithms to maintain revenue constraint.

Stochastic Conservative Bandits

- $K+1$ actions or *arms*, each with mean reward $\mu_i \in [0,1]$ for $i \in \{0,1,\dots,K\}$. "Default" action is $i=0$; μ_0 is known and other μ_i are unknown.
- Learner chooses action I_t at round t and receives reward $X_t = \mu_{I_t} + \eta_t$, where η_t is sub-gaussian noise.
- With high probability $(1-\delta)$, must satisfy constraint

$$\sum_{t=1}^{n} \mu_{I_t} \geq (1-\alpha)\mu_0 n, \qquad \text{for all } n;$$

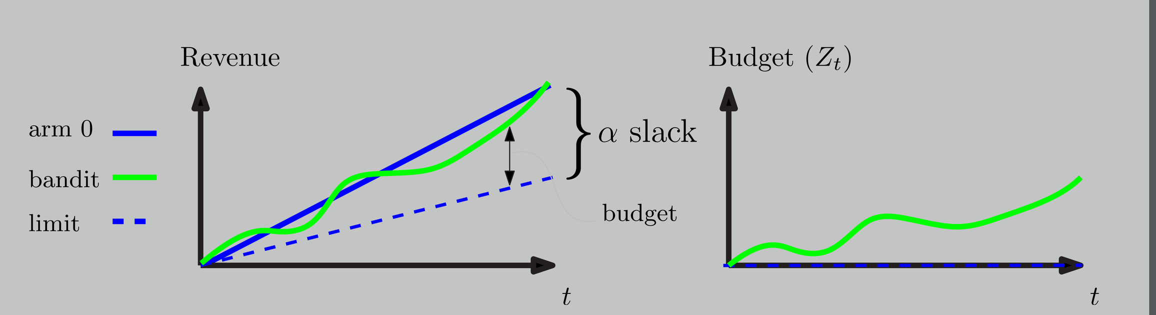
- You choose α and δ (e.g. $\alpha = 0.1$ loses up to 10% revenue compared to the default action).

(Pseudo) regret of learner: gap between reward and maximum achievable in hindsight (by always choosing best action):

$$R_n = \sum_{t=1}^{n} (\max_i \mu_i - \mu_{I_t}).$$

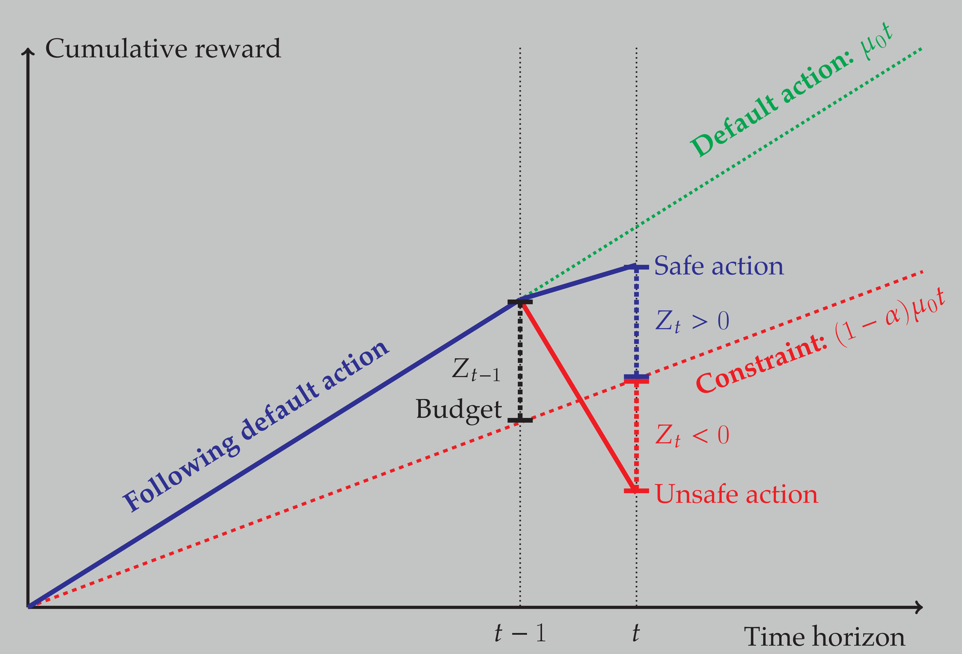
The Challenge: Minimizing regret requires exploration to find best arm, but maintaining constraint requires choosing default arm very often.

Budget



Budget: $Z_t = \sum_{s=1}^{t} \mu_{I_s} - (1-\alpha)t\mu_0.$

- Constraint satisfied iff budget is positive.
- Default action is *safe*: it *increases* budget by $\alpha\mu_0$.
- Can use high probability lower bounds for unknown μ_i to bound budget.
- **Figure:** Learner chooses default arm up to round $t-1$, accumulating budget Z_{t-1}. Then it can choose a safe arm (blue) keeping $Z_t > 0$, but an unsafe arm (red) would make $Z_t < 0$.



- **Conservative UCB:** choose arm with greatest UCB, unless doing so would make the budget's LCB negative.

Conservative UCB Algorithm

Input: K, μ_0, δ, $\psi^\delta(\cdot)$
for $t \in 1,2,\dots$ **do**
  ▷ Compute confidence intervals…
  $\theta_0(t), \lambda_0(t) \leftarrow \mu_0$ ▷ …for known μ_0,
  for $i \in 1,\dots,K$ **do** ▷ …for other arms,
    $\Delta_i(t) \leftarrow \sqrt{\psi^\delta(T_i(t-1))/T_i(t-1)}$
    $\theta_i(t) \leftarrow \hat{\mu}_i(t-1) + \Delta_i(t)$
    $\lambda_i(t) \leftarrow \max\{0, \hat{\mu}_i(t-1) - \Delta_i(t)\}$
  $J_t \leftarrow \arg\max_i \theta_i(t)$ ▷ …and find UCB arm.
  ▷ Compute budget and…
  $\xi_t \leftarrow \sum_{s=1}^{t-1} \lambda_{I_s}(t) + \lambda_{J_t}(t) - (1-\alpha)t\mu_0$
  if $\xi_t \geq 0$ **then**
    $I_t \leftarrow J_t$ ▷ …choose UCB arm if safe,
  else
    $I_t \leftarrow 0$ ▷ …default arm otherwise.

$\psi(n) \approx \log\log n$ is inspired by a concentration inequality. A good choice is in the paper.

Variants of Algorithm

- Unknown μ_0 (learn it by taking default action)
- Expected regret/budget (instead of high probability)

Upper Bound on Regret

Theorem: For all rounds n, Conservative UCB satisfies the following with probability at least $1-\delta$:

- Minimum reward: $\sum_{t=1}^{n} \mu_{I_t} \geq (1-\alpha)n\mu_0$,
- Maximum regret: $R_n \leq O\left(\sqrt{nKL} + KL/\alpha\mu_0\right)$, where $L = \psi^\delta(n) \approx \log\log n$

Lower Bound on Regret

Theorem: There are "hard" environments: any algorithm satisfying constraint must have regret

$$E_\mu[R_n] \geq \Omega(\sqrt{nK} + K/\alpha\mu_0).$$

- Can specify number of arms K, rounds n, and reward of default arm μ_0 (sufficiently far from 0 and 1).
- Almost matches Conservative UCB regret.

Adversarial Conservative Bandits

Adversary generates rewards $X_{t,i} \in [0,1]$ (at round t for arm $i \neq 0$), while $X_{t,0}$ is held constant. Constraint is:

$$\sum_{t=1}^{n} X_{t,I_t} \geq (1-\alpha)\sum_{t=1}^{n} X_{t,0}$$

Safe-play strategy: Act according to "base" any-time high-probability adversarial bandit algorithm (e.g. Exp3-IX of Neu, 2015) when safe. Otherwise, default action.

Theorem: Let $t_0 = \max\{t \geq 1 \mid \alpha\mu_0 t \leq R_t^\delta + \mu_0\}$. When the base algorithm is $\{R_t^\delta\}$ admissible w.p. $1-\delta$ for any n, safe-play satisfies budget constraint while achieving regret $R_n \leq t_0 + R_n^\delta$.

Corollary: Safe-play strategy applied to Exp3-IX gives w.p. $1-\delta$ (where $L \approx \log n$)

$$R_n \leq O\left(\sqrt{Kn\log K} + KL^2/\alpha^2\mu_0^2\right).$$

- Maintaining constraint costs more regret here $(KL^2/\alpha^2\mu_0^2)$ than in stochastic case $(KL/\alpha\mu_0)$. *Can we do better?*

Experiments

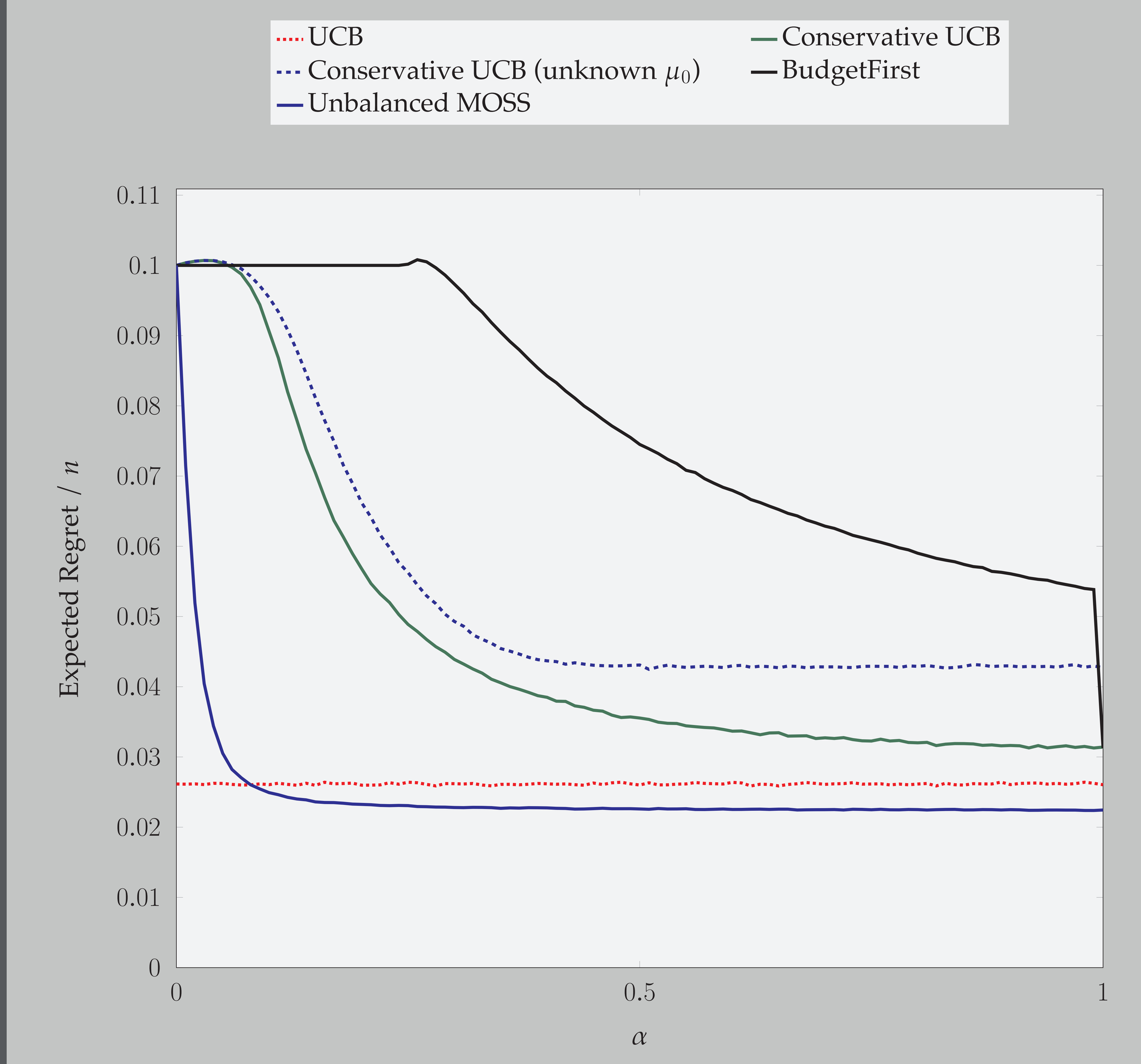
Environment:
$K = 5$ arms; $\mu_0 = 0.5$, $\mu_1 = 0.6$, $\mu_2 = \mu_3 = \mu_4 = 0.4$.

Comparing following algorithms:

Algorithm	Constraint?	Unknown μ_0?
UCB	×	✓
Unbalanced MOSS	✓ (at end)	×
Budget-First	✓	×
Conservative UCB	✓	✓ (optional)

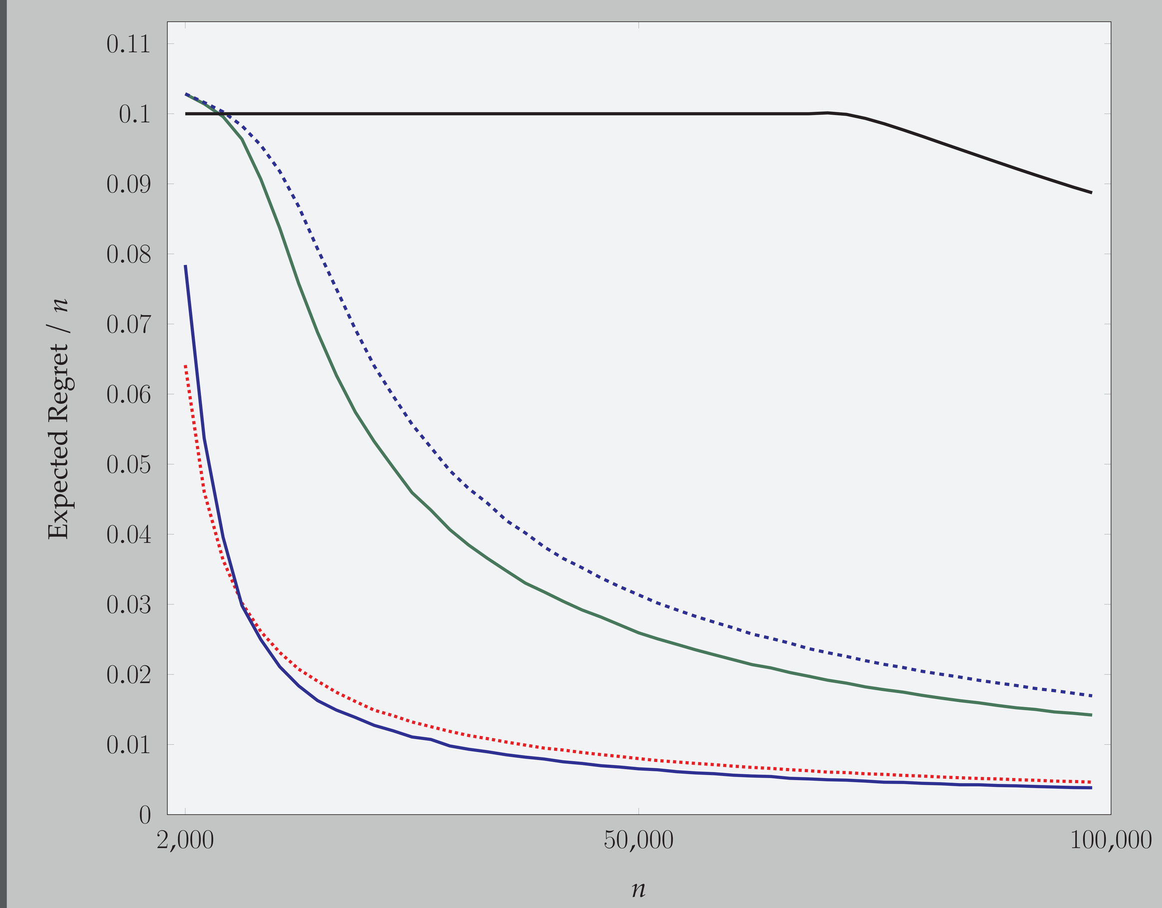
First experiment:

- Regret after $n = 10^4$ steps with probability $\delta = 1/n$
- Varying constraint harshness (α)



Second experiment:

- Varying time horizon n with probability $\delta = 1/n$
- Fixed $\alpha = 0.1$



Discussion

- Conservative UCB pays price for maintaining constraint, getting worse as α becomes small
- Eventually almost as good as UCB
- Small advantage to know μ_0, even when unconstrained ($\alpha = 1$)
- Unbalanced MOSS: better performance but only satisfies constraint at end; no high-probability bounds

Summary

- Introduced a new multi-armed bandit setting: actual return must be close to that of a default action *uniformly in time*
- Conservative UCB algorithm (and variants) for stochastic problems; Safe-Play strategy for adversarial
- Conservative UCB: near-optimal
- Gap between lower and upper bound for adversarial case