A simple MAB algorithm with optimal variation bounded regret

Elad Hazan & Satyen Kale

@ Technion & Yahoo/IBM
Multi-Armed Bandits: [Robbins ’52] a fundamental decision problem

Decision maker chooses an arm \( x_t \in [n] \)

Observe loss \( l_t(x_t) \in [0,1] \)

And repeat...

\[
\text{Average Regret} = \frac{\sum_t l_t(x_t)}{T} - \min_x \frac{\sum_t l_t(x)}{T} \rightarrow 0
\]
State Of the art

1. Stochastic setting:

\[ \text{Regret} = O(\log T) \] [Auer, Cesa-Bianchi, Fischer]

2. Non-stochastic MAB [Auer, Cesa-Bianchi, Freund, Schapire], [Audibert-Bubeck]

\[ \text{Regret} = O(\sqrt{Tk}) \]

Both results are optimal (in stochastic / worst case settings respec.)
Can we be more optimal?

Consider adversarial setting.

Tighter measure of regret:

\[
\text{Variation} = Q = \sum_t \| l_t - \mu \|^2, \quad \mu = \frac{1}{T} \sum_t l_t
\]

- Natural measure, variance in statistical setting, always < T.

- [Cesa-Bianchi, Mansour, Stoltz] Can we get regret bounded by \( O(\sqrt{Q}) \)?

(this is bounded by \( O(\sqrt{Tk}) \))
Known Variational bounds / bandits

[HK’08, HK’09] – variational bounds for full information OLO & exp-concave (portfolio selection)

\[ \text{Regret} = O(\sqrt{Q}), O(\log Q) \]

[HK ’09]: \[ \text{Regret} = O(k^2 \sqrt{Q \log T}) \]

Complicated algorithm, based on [Abernethy, Hazan, Rakhlin] alg. for OLO.

Tools:
- Reservoir sampling to estimate mean cost
- Deviation estimators – according to historical mean
The Question

Does there exist a (simple) algorithm with regret bounded by:

\[ \text{Regret} = O(\sqrt{Q}) \]

Alg should look like EXP3, with additional tricks (reservoir sampling, history-adjusted estimators)