Improved Algorithms for Linear Stochastic Bandits

Yasin Abbasi-Yadkori, Dávid Pál, Csaba Szepesvári

Department of Computing Science, University of Alberta
&
Google, New York
abbasiya@ualberta.ca, dpal@google.com, szepesva@ualberta.ca

December 14, 2011

Linear Bandits

In round $t = 1, 2, \ldots$

- Choose an action X_t from a set $D_t \subset \mathbb{R}^d$.
- Receive a reward $Y_t = \langle X_t, \theta_* \rangle + \eta_t$
- Goal: Maximize total reward.
- Web advertisement, online shortest path, ...

- Weights θ_* are unknown but fixed. $\|\theta_*\|_2 \leq S$.
- Noise is conditionally R-sub-Gaussian i.e.

$$\forall \gamma \in \mathbb{R} \qquad \mathbf{E}[e^{\gamma \eta_t} \mid X_{1:t}, \eta_{1:t-1}] \leq \exp\left(\frac{\gamma^2 R^2}{2}\right) .$$

Regret of the Bandit Algorithm

- Optimism in the Face of Uncertainty.

- Build a confidence set. With probability $\geq 1-\delta$,

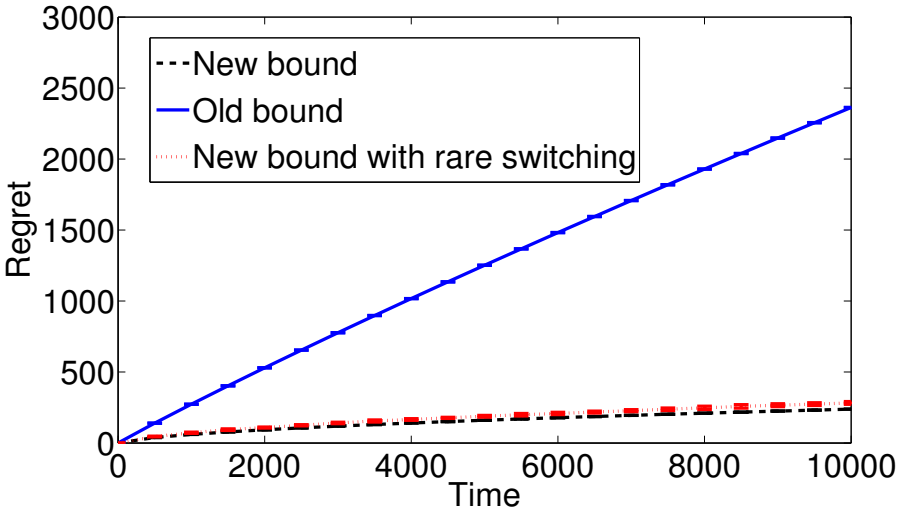
$$\|\widehat{\theta}_t - \theta_*\|_{V_t} \leq R\sqrt{2\ln\left(\frac{\det(V_t)^{1/2}}{\delta\det(\lambda I)^{1/2}}\right)} + S\sqrt{\lambda}$$

  A tight data dependent confidence set.
  ($V_t = \sum_{s=1}^{t} X_t X_t^\top$, $\widehat{\theta}_t$: ridge-regression solution, λ: regularizer.)

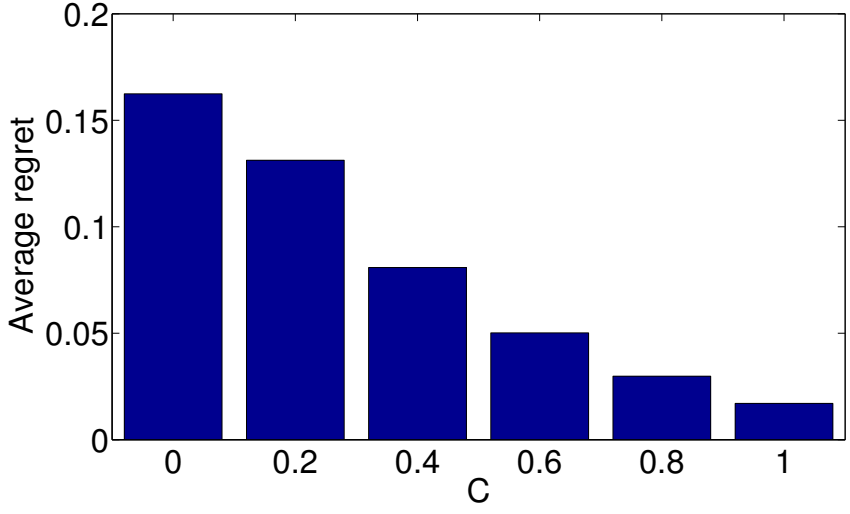
- Improvements over previous results (Dani et al., 2008):

  Worst-case regret:
  $O(d\log(n)\sqrt{n\log(n/\delta)}) \rightarrow O(d\log(n)\sqrt{n} + \sqrt{dn\log(n/\delta)})$

  Problem-dependent regret with "gap" Δ :
  $O(\frac{d^2}{\Delta}\log(n/\delta)\log^2(n)) \rightarrow O(\frac{\log 1/\delta}{\Delta}(\log^2(n) + d^2 + d\log n))$



Vast improvements
compared to Dani et al. (2008)



Computational Improvements

See our poster W082!