Improved Algorithms for Linear Stochastic Bandits (Abbasi-Yadkori, Pál, and Szepesvári 2011)

Charlie Godfrey, Oliver Knitter, Kapila Kottegoda, Yunpeng Shi

January 20, 2020

Decisions and Rewards

- At each time $t \in \mathbb{N}$, there's a **decision set** $D_t \subset \mathbb{R}^d$, and

Decisions and Rewards

- At each time $t \in \mathbb{N}$, there's a **decision set** $D_t \subset \mathbb{R}^d$, and
- player must choose an **action** $X_t \in D_t$.

Decisions and Rewards

- At each time $t \in \mathbb{N}$, there's a **decision set** $D_t \subset \mathbb{R}^d$, and
- player must choose an **action** $X_t \in D_t$.
- The **reward** is

$$Y_t = \langle X_t, \theta_* \rangle + \eta_t \in \mathbb{R}$$

Decisions and Rewards

- At each time $t \in \mathbb{N}$, there's a **decision set** $D_t \subset \mathbb{R}^d$, and
- player must choose an **action** $X_t \in D_t$.
- The **reward** is

$$Y_t = \langle X_t, \theta_* \rangle + \eta_t \in \mathbb{R}$$

- Here η_t is a 1 sub-Gaussian random variable (noise) such that

$$\mathrm{E}[\eta_t | X_{1:t}, \eta_{1:t-1}] = 0$$

Decisions and Rewards

- At each time $t \in \mathbb{N}$, there's a **decision set** $D_t \subset \mathbb{R}^d$, and
- player must choose an **action** $X_t \in D_t$.
- The **reward** is

$$Y_t = \langle X_t, \theta_* \rangle + \eta_t \in \mathbb{R}$$

- Here η_t is a 1 sub-Gaussian random variable (noise) such that

$$\mathrm{E}[\eta_t | X_{1:t}, \eta_{1:t-1}] = 0$$

- θ_* is an unknown vector in $\mathbb{R}^d$, which the player needs to estimate.

Reward and Regret

reward at time t	$\mathrm{E}[\sum_{s\leq t} Y_s] = \sum_{s\leq t}\langle X_s, \theta_*\rangle$
optimal strategy	$x_t^* = \arg\max_{x\in D_t}\langle x, \theta_*\rangle$
max possible reward at time t	$\sum_{s\leq t}\langle x_s^*, \theta_*\rangle$
(pseudo-)regret at time t	$R_t = \sum_{s\leq t}\langle x_s^* - X_s, \theta_*\rangle$

Optimism in the Face of Uncertainty (Linearized)

input: decision sets D_t, initial confidence set C_0
loop
$t = 1, 2, 3, \ldots$
$(X_t, \tilde{\theta}_t) = \arg\max_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle$
play X_t, **observe** Y_t
update C_t
end loop

Confidence Ellipsoids

- confidence sets C_t are updated to **ellipsoids** computed via **ridge regression** on the (X_s, Y_s) $(s \leq t)$

Confidence Ellipsoids

- confidence sets C_t are updated to **ellipsoids** computed via **ridge regression** on the (X_s, Y_s) $(s \leq t)$
- set $\bar{V}_t = \lambda I + \sum_{s \leq t} X_s X_s^T$ and $\hat{\theta}_t = \bar{V}_t^{-1} \sum_{s \leq t} Y_s X_s$

Confidence Ellipsoids

- confidence sets C_t are updated to **ellipsoids** computed via **ridge regression** on the (X_s, Y_s) $(s \leq t)$
- set $\bar{V}_t = \lambda I + \sum_{s \leq t} X_s X_s^T$ and $\hat{\theta}_t = \bar{V}_t^{-1} \sum_{s \leq t} Y_s X_s$
- then

$$C_t = \{\theta \in \mathbb{R}^d \mid \|\theta - \hat{\theta}_t\|_{\bar{V}_t} \leq \epsilon_t\}$$

where

$$\epsilon_t = \sqrt{2 \log\left(\frac{\det(\bar{V}_t)^{\frac{1}{2}} \det(\lambda I)^{-\frac{1}{2}}}{\delta}\right)} + \lambda^{\frac{1}{2}} S$$

.

Regret bound of the OFUL algorithm

Theorem (Abbasi-Yadkori, Pál, and Szepesvári 2011)

Assume that $\|X_t\| \le L$ *and* $\|\theta_*\| \le S$, *for some constants* $L, S > 0$, *and that* $\langle x, \theta_* \rangle \in [-1, 1]$ *for all* $x \in D_t$. *Suppose* $\lambda \ge 1$. *Then with probability at least* $1 - \delta$ *the OFUL algorithm achieves*

$$R_t \le 4\sqrt{td\log(\lambda + \frac{tL}{d})}(\sqrt{\lambda}S + \sqrt{2\log(\frac{1}{\delta}) + d\log(1 + \frac{tL}{\lambda d})})$$

Comparison of Regret Bounds

	Dani et al. 08 [1]	This Work
Linear Bandits	$O\left(d\log(t)\sqrt{t\log(\frac{t}{\delta})}\right)$	$O\left(d\log(t)\sqrt{t}+\sqrt{dt\log(\frac{t}{\delta})}\right)$
d-armed	$O(d\log(t)/\Delta)$	$O(d\log(\frac{1}{\delta})/\Delta)$
Problem Dependent	$O(\frac{d^2}{\Delta}\log(\frac{t}{\delta})\log^2(t))$	$O(\frac{\log(1/\delta)}{\Delta}(\log(t)+d\log\log(t))^2)$

[1] (Dani, Hayes, and Kakade 2008)

A Self-Normalized Bound for Vector-Valued Martingales

Theorem (Abbasi-Yadkori, Pál, and Szepesvári 2011)

Let

- $F_t = \sigma(X_1, \ldots, X_{t+1}, \eta_1, \ldots, \eta_t)$, *so* $\{F_t\}_{t=0}^{\infty}$ *is a filtration of* σ*-algebras*
- η_t *be* 1*-sub-Gaussian conditioned on* F_{t-1}, *with* $\mathrm{E}[\eta_t \,|\, F_{t-1}] = 0$
- $\overline{V}_t = V_t + \lambda I, \qquad V_t = \sum_{s=1}^{t} X_s X_s^\top \qquad S_t = \sum_{s=1}^{t} \eta_s X_s.$

Then, for any $\delta > 0$, *with probability at least* $1 - \delta$

$$\|S_t\|^2_{\overline{V}_t^{-1}} \leq 2 \log \left(\frac{\det(\overline{V}_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)$$

for all $t \geq 0$.

Proof Ideas

Lemma

In the setting of the theorem, let τ be any stopping time w.r.t $\{F_t\}_{t=0}^{\infty}$, then for any $\delta > 0$, with probability $1 - \delta$

$$\|S_t\|^2_{\overline{V}_\tau^{-1}} \leq 2\log\left(\frac{\det(\overline{V}_\tau)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)$$

- $B_t(\delta) = \left\{ w \in \Omega : \|S_t\|^2_{\overline{V}_t^{-1}} > 2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)\right\}$
- τ = the first time that a bad event $B_t(\delta)$ occurs
- τ is a stopping time

- $B_t(\delta) = \left\{ w \in \Omega : \|S_t\|^2_{\overline{V}_t^{-1}} > 2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)\right\}$
- τ = the first time that a bad event $B_t(\delta)$ occurs
- τ is a stopping time

$$
\begin{aligned}
\Pr\left(\bigcup_{t\geq 0} B_t(\delta)\right) &= \Pr(\tau < \infty) \\
&\leq \Pr\left(\|S_\tau\|^2_{\overline{V}_\tau^{-1}} > 2\log\left(\frac{\det(\overline{V}_\tau)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)\right) \\
&\leq \delta
\end{aligned}
$$

Conclusion

- OFUL is a UCB-type algorithm for **linear** bandits, with an $\mathcal{O}(d\log(t)\sqrt{t}+\sqrt{dt\log(\frac{t}{\delta})})$ regret bound.
- Self-normalized martingale bound has many further applications, e.g. to finite-time identification of linear dynamical systems (Sarkar and Rakhlin 2018).

References

Abbasi-Yadkori, Yasin, Dávid Pál, and Csaba Szepesvári (2011). "Improved Algorithms for Linear Stochastic Bandits". In: *NIPS*.

Dani, Varsha, Thomas P. Hayes, and Sham M. Kakade (2008). "Stochastic Linear Optimization under Bandit Feedback". In: *COLT*.

Sarkar, Tuhin and Alexander Rakhlin (2018). *Near optimal finite time identification of arbitrary linear dynamical systems*. arXiv: 1812.01251v7 [cs.SY].