Improved Algorithms for Linear Stochastic Bandits

일 시 : 2021년 7월 21일
장 소 : SDM LAB.

Intern, An Byeongwoo

0. Abstract

- Stochastic multi-armed bandit problem과 linear stochastic multi-armed bandit problem에 대한 알고리즘의 이론적 분석과 실증적인 성능 개선하고자 함.
- 특히, Auer's UCB 알고리즘의 수정을 통해 high probability constant regret을 달성함을 보여줌.
- logarithmic factor를 통한 regret bound를 개선한다.
- 새로운 vector-valued martingale을 위한 꼬리 부등식을 사용하여 더 작은 신뢰 집합을 만든다.

1. Introduction

- Linear stochastic bandit problem은 순차적 의사결정 문제.
- n번의 시행동안 가능한 많은 reward를 받는 것이 목표.
- 여러 변형모델이 존재하고, 모두 optimism in the face of uncertainty(OFU) principle에 기반한다.
- OFU principle은 exploration-exploitation 딜레마를 해결할 수 있다.
- 기본 개념은 선형 함수 계수 벡터의 신뢰 집합의 유지하는 것이다.
- 매 라운드, 알고리즘은 신뢰 집합에서 예측치를 선택하고 예상 reward가 최대가 되는 행동을 한다.
- 따라서, 문제를 과거에 관측된 action-reward에 기반한 선형 함수의 계수 벡터에 대한 신뢰 집합의 구성으로 축소시킬 수 있다.
- 미래의 행동은 과거의 행동으로부터 독립적이지 않기 때문에 해결하기 쉬운 문제가 아니다.
- 몇몇 논문은 이러한 문제를 간과했다.
- 올바른 해결법을 본 논문에서 새로운 마틴게일 기법으로 제시한다.

1. Introduction

- 작은 신뢰 집합을 가질수록, 더 나은 regret bound를 얻을 수 있고, 알고리즘의 성능이 더 좋다.
- 소개할 알고리즘을 통해, 신뢰 집합의 크기를 줄일 수 있다.
- First, 매 단계에서 신뢰 집합이 균일하게 유효하다. (union bound를 피함으로써 log(n)만큼 줄일 수 있다)
- Second, 항상, 대체로 경험적인 양으로 대체된다는 점에서 '더 실증적'이다.
- 신뢰 집합을 만들기 구성하기 위해, 새로운 martingale tail ineqaulity를 증명했다.
- 새로운 신뢰 집합을 이용하여 UCB 알고리즘을 수정했고, δ를 input으로 받기 때문에 regret은 n에 종속되지 않고, δ의 종속된다.
- δ=1/n 일 때, 새 알고리즘은 같은 expected regret bound($O((d\log n$ / Δ)를 가진다.

1. Introduction

- CONFIDENCBALL 알고리즘은 확률이 적어도 1- δ 일 때, regret이 최대 $O(d\log(n)\sqrt{n\log(n/\delta)})$ 라 했지만 수정된 알고리즘은 최대 $O(d\log(n)\sqrt{n} + \sqrt{dn\log(n/\delta)})$ 임을 밝혔다.
- 또한, dependent regret bound $O(\frac{d^2}{\Delta}\log(n/\delta)\log^2(n))$ 에 대해, 향상된 $O(\frac{\log(1/\delta)}{\Delta}(\log(n) + d\log\log n)^2)$ bound증명했다.

1.2 The Learning Model

- 매 라운드(t)마다 learner에게 행동 X_t을 선택할 수 있는 유클리드 공간에 속하는 decision set D_t이 주어진다.
- 그 후, learner는 reward $Y_t = \langle X_t, \theta_* \rangle + \eta_t$ where $\theta_* \in \mathbb{R}^d$ 를 관측할 수 있다.
- θ_* 는 unknown parameter η_t는 $\mathbf{E}[\eta_t \mid X_{1:t}, \eta_{1:t-1}] = 0$ 를 만족하는 random noise
- learner의 목표는 그의 total reward $\sum_{t=1}^{n} \langle X_t, \theta_* \rangle$ 를 최대화 하는 것.
- θ_* 를 알고 있다면, t시점의 최적 전략은 $x_t^* = \mathrm{argmax}_{x \in D_t} \langle x, \theta_* \rangle$
- 최적 전략의 total reward와 leaner가 얻은 total reward의 차이를 pseudo-regret이라 한다.

$$R_n = \left(\sum_{t=1}^{n} \langle x_t^*, \theta_* \rangle \right) - \left(\sum_{t=1}^{n} \langle X_t, \theta_* \rangle \right) = \sum_{t=1}^{n} \langle x_t^* - X_t, \theta_* \rangle$$

ex) $\vec{x}_1^* \cdot \vec{\theta}_* - \vec{X}_1 \vec{\theta}_*$

$= \overrightarrow{(x_1^* - X_1)}\, \vec{\theta}_*$

$= \langle x_1^* - X_1, \theta_* \rangle$

- 알고리즘의 목표는 regret R_n을 최소화 하는 것.

1.2 The Learning Model

- Regret의 의미있는 upper bound를 얻기 위한 가정

- Dt는 유계집합이다.
- η_t는 조건부 R-sub-Gaussian 이다. R>=0인 고정 상수 일 때.

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

Definition 1.2. A random variable $X \in \mathbb{R}$ is said to be *sub-Gaussian* with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and its moment generating function satisfies

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right), \quad \forall s \in \mathbb{R}. \tag{1.2}$$

1.2 The Learning Model

- $\mathbf{E}[\eta_t \mid X_{1:t}, \eta_{1:t-1}] = 0.$ 와 $\mathbf{Var}[\eta_t \mid F_t] \leq R^2$ 임을 의미.

$$M_{\eta_t}(\lambda) \leq \exp\left(\frac{\lambda^2 R^2}{2}\right)$$

$$M_{\eta_t}(\lambda)' \leq R^2\lambda \cdot e^{\frac{\lambda^2 R^2}{2}}$$

$$M_{\eta_t}(\lambda)'' \leq R^2 \cdot e^{\frac{\lambda^2 R^2}{2}} + (R^2\lambda)^2 e^{\frac{\lambda^2 R^2}{2}}$$

$$M_{\eta_t}(0)'' - M_{\eta_t}(0)' \leq R^2$$

$$\therefore V(\eta_t \mid F_t) \leq R^2$$

2. Optimism in the Face of Uncertainty

- θ_* 를 위해 $C_{t-1} \subseteq \mathbb{R}^d$ 를 유지하는 것이 기본 개념.

$$(X_t, \widetilde{\theta}_t) = \underset{(x,\theta)\in D_t \times C_{t-1}}{\operatorname{argmax}} \langle x, \theta \rangle$$

- 문제의 핵심은 신뢰 집합 Ct의 구성.

3. Self-Normalized Tail Inequality for Vetor-Valued Martingales

- $\{D_t\}_{t=1}^{\infty}$ 가 임의적이라면, $X_t \in D_t$ 또한 임의적.
- 다루기 힘든 복잡한 통계적인 종속을 가진 sequence $\{X_t\}_{t=1}^{\infty}$ 를 만듦.
- 따라서 신뢰 집합을 도출함에 있어서 $\{X_t\}_{t=1}^{\infty}$ 에 대한 어떠한 가정도 하지 않는 것이 좋다.
- $\{S_t\}_{t=0}^{\infty}$ 은 θ_*를 위한 신뢰 집합을 만드는데 중요한 $\{F_t\}_{t=0}^{\infty}$ 에 관련된 마틴게일.
- Theorem1은 높은 확률로 마틴게일이 0에 근접함을 보인다.

Theorem 1

Theorem 1 (Self-Normalized Bound for Vector-Valued Martingales). *Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is F_t-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$ i.e.*

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

Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is F_{t-1}-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$, define

$$\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top \qquad\qquad S_t = \sum_{s=1}^{t} \eta_s X_s .$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,

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

Note that the deviation of the martingale $\|S_t\|^2_{\overline{V}_t^{-1}}$ is measured by the norm weighted by the matrix $\overline{V}_t^{-1}$ which is itself derived from the martingale, hence the name "self-normalized bound".

Theorem 1

Lemma 8. *Let $\lambda \in \mathbb{R}^d$ be arbitrary and consider for any $t \geq 0$*

$$M_t^\lambda = \exp\left(\sum_{s=1}^{t}\left[\frac{\eta_s \langle \lambda, X_s \rangle}{R} - \frac{1}{2}\langle \lambda, X_s \rangle^2\right]\right) .$$

Let τ be a stopping time with respect to the filtration $\{F_t\}_{t=0}^{\infty}$. Then M_τ^λ is almost surely well-defined and

$$\mathbf{E}[M_\tau^\lambda] \leq 1 .$$

Proof of Lemma 8. We claim that $\{M_t^\lambda\}_{t=0}^{\infty}$ is a supermartingale. Let

$$D_t^\lambda = \exp\left(\frac{\eta_t \langle \lambda, X_t \rangle}{R} - \frac{1}{2}\langle \lambda, X_t \rangle^2\right) .$$

Observe that by conditional R-sub-Gaussianity of η_t we have $\mathbf{E}[D_t^\lambda \mid F_{t-1}] \leq 1$. Clearly, D_t^λ is F_t-measurable, as is M_t^λ. Further,

$$\mathbf{E}[M_t^\lambda \mid F_{t-1}] = \mathbf{E}[M_1^\lambda \cdots D_{t-1}^\lambda D_t^\lambda \mid F_{t-1}] = D_1^\lambda \cdots D_{t-1}^\lambda \, \mathbf{E}[D_t^\lambda \mid F_{t-1}] \leq M_{t-1}^\lambda ,$$

showing that $\{M_t^\lambda\}_{t=0}^{\infty}$ is indeed a supermartingale and in fact $\mathbf{E}[M_t^\lambda] \leq 1$.

Now, we argue that M_τ^λ is well-defined. By the convergence theorem for nonnegative supermartingales, $M_\infty^\lambda = \lim_{t\to\infty} M_t^\lambda$ is almost surely well-defined. Hence, M_τ^λ is indeed well-defined independently of whether $\tau < \infty$ holds or not. Next, we show that $\mathbf{E}[M_\tau^\lambda] \leq 1$. For this let $Q_t^\lambda = M_{\min\{\tau,t\}}^\lambda$ be a stopped version of $(M_t^\lambda)_t$. By Fatou's Lemma, $\mathbf{E}[M_\tau^\lambda] = \mathbf{E}[\liminf_{t\to\infty} Q_t^\lambda] \leq \liminf_{t\to\infty} \mathbf{E}[Q_t^\lambda] \leq 1$, showing that $\mathbf{E}[M_\tau^\lambda] \leq 1$ indeed holds. □

$$M_t^\lambda = \exp\left(\sum_{s=1}^{t}\left[\frac{\eta_s\langle\lambda, X_s\rangle}{R} - \frac{1}{2}\langle\lambda, X_s\rangle^2\right]\right)$$

$$E[M_\tau^\lambda] \leq 1$$

$$D_t^\lambda = \exp\left(\frac{\eta_t\langle\lambda, X_t\rangle}{R} - \frac{1}{2}\langle\lambda, X_t\rangle^2\right)$$

$E[D_t^\lambda \mid F_{t-1}] \leq 1$

$E[M_t^\lambda \mid F_{t-1}] = E[M_1^\lambda \cdots D_{t-1}^\lambda \cdot D_t^\lambda \mid F_{t-1}]$

$$\Rightarrow M_t^\lambda = \exp\left(\left(\frac{\eta_1\langle\lambda, X_1\rangle}{R} - \frac{1}{2}\langle\lambda, X_1\rangle^2\right) + \left(\frac{\eta_2\langle\lambda, X_2\rangle}{R} - \frac{1}{2}\langle\lambda, X_2\rangle^2\right)\right.$$

$$\left. + \cdots + \left(\frac{\eta_{t-1}\langle\lambda, X_{t-1}\rangle}{R} - \frac{1}{2}\langle\lambda, X_{t-1}\rangle^2\right) + \left(\frac{\eta_t\langle\lambda, X_t\rangle}{R} - \frac{1}{2}\langle\lambda, X_t\rangle^2\right)\right)$$

$$= M_1^\lambda \cdots D_{t-1}^\lambda \cdot D_t^\lambda$$

$= D_1^\lambda \cdots D_{t-1}^\lambda \, E[D_t^\lambda \mid F_{t-1}] \leq M_{t-1}^\lambda$

$\Rightarrow E[D_t^\lambda \mid F_{t-1}] \leq 1$

$D_1^\lambda \cdots D_{t-1}^\lambda \, E[D_t^\lambda \mid F_{t-1}] \leq D_1^\lambda \cdots D_{t-1}^\lambda = M_{t-1}^\lambda$

Theorem 1

Lemma 9 (Self-normalized bound for vector-valued martingales). *Let τ be a stopping time with respect to the filtration $\{F_t\}_{t=0}^{\infty}$. Then, for $\delta > 0$, with probability $1-\delta$,*

$$\|S_\tau\|^2_{\overline{V}_\tau^{-1}} \le 2R^2 \log\left(\frac{\det(\overline{V}_\tau)^{1/2}\det(V)^{-1/2}}{\delta}\right).$$

$$M_t^\lambda = \exp\left(\overset{①}{\langle \lambda, S_t\rangle} - \frac{1}{2}\overset{②}{\|\lambda\|^2_{V_t}}\right) \qquad V_t = \sum_{s=1}^{t} X_s X_s^T$$

① $\sum_{s=1}^{t} \eta_s \langle \lambda, X_s\rangle = \langle \lambda, \sum_{s=1}^{t} \eta_s X_s\rangle = \langle \lambda, S_t\rangle$ $\qquad S_t = \sum_{s=1}^{t} \eta_s X_s$

② $\|\lambda\|^2_{V_t} = \lambda^T V_t \lambda = \lambda^T\left(\sum_{s=1}^{t} X_s X_s^T\right)\lambda = \langle \lambda, X_s\rangle^2$

ex) $\lambda = \begin{pmatrix} a \\ b \end{pmatrix}$, $X_s = \begin{pmatrix} x \\ y \end{pmatrix}$ $(s=1)$

$X_1 X_1^T = \begin{pmatrix} x \\ y \end{pmatrix}(x \ y) = \begin{pmatrix} x^2 & xy \\ xy & y^2 \end{pmatrix}$

$(a \ b)\begin{pmatrix} x^2 & xy \\ xy & y^2 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = a^2x^2 + 2abxy + b^2y^2 = (ax+by)^2 = \langle \lambda, X_1\rangle^2$

Proof of Lemma 9. Without loss of generality, assume that $R = 1$ (by appropriately scaling S_t, this can always be achieved). Let

$$V_t = \sum_{s=1}^{t} X_s X_s^\top \qquad M_t^\lambda = \exp\left(\langle \lambda, S_t\rangle - \frac{1}{2}\|\lambda\|^2_{V_t}\right) \qquad M_t^\lambda = \exp\left(\sum_{s=1}^{t}\left[\frac{\eta_s\langle \lambda, X_s\rangle}{R} - \frac{1}{2}\langle \lambda, X_s\rangle^2\right]\right)$$

Notice that by Lemma 8, the mean of M_τ^λ is not larger than one.

Let Λ be a Gaussian random variable which is independent of all the other random variables and whose covariance is V^{-1}. Define

$$M_t = \mathbf{E}[M_t^\Lambda \mid F_\infty],$$

where F_∞ is the tail σ-algebra of the filtration i.e. the σ-algebra generated by the union of the all events in the filtration. Clearly, we still have $\mathbf{E}[M_\tau] = \mathbf{E}[\mathbf{E}[M_\tau^\Lambda \mid \Lambda]] \le 1$.

law of total expectation.

Theorem 1

Let us calculate M_t. Let f denote the density of Λ and for a positive definite matrix P le $c(P) = \sqrt{(2\pi)^d / \det(P)} = \int \exp(-\frac{1}{2}x^\top Px)dx$. Then,

$$\begin{aligned}
M_t &= \int_{\mathbb{R}^d} \exp\left(\langle \lambda, S_t\rangle - \frac{1}{2}\|\lambda\|^2_{V_t}\right) f(\lambda)\, d\lambda \\
&= \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|\lambda - V_t^{-1}S_t\|^2_{V_t} + \frac{1}{2}\|S_t\|^2_{V_t^{-1}}\right) f(\lambda)\, d\lambda \\
&= \frac{1}{c(V)} \exp\left(\tfrac{1}{2}\|S_t\|^2_{V_t^{-1}}\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\{\|\lambda - V_t^{-1}S_t\|^2_{V_t} + \|\lambda\|^2_V\right\}\right) d\lambda\, .
\end{aligned}$$

Elementary calculation shows that if P is positive semi-definite and Q is positive definite

$$\|x - a\|^2_P + \|x\|^2_Q = \left\|x - (P+Q)^{-1}Pa\right\|^2_{P+Q} + \|a\|^2_P - \|Pa\|^2_{(P+Q)^{-1}}\, .$$

Therefore,

$$\begin{aligned}
\|\lambda - V_t^{-1}S_t\|^2_{V_t} + \|\lambda\|^2_V &= \|\lambda - (V+V_t)^{-1}S_t\|^2_{V+V_t} + \|V_t^{-1}S_t\|^2_{V_t} - \|S_t\|^2_{(V+V_t)^{-1}} \\
&= \|\lambda - (V+V_t)^{-1}S_t\|^2_{V+V_t} + \|S_t\|^2_{V_t^{-1}} - \|S_t\|^2_{(V+V_t)^{-1}}\, ,
\end{aligned}$$

which gives

$$\begin{aligned}
M_t &= \frac{1}{c(V)} \exp\left(\frac{1}{2}\|S_t\|^2_{(V+V_t)^{-1}}\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|\lambda - (V+V_t)^{-1}S_t\|^2_{V+V_t}\right) d\lambda \\
&= \frac{c(V+V_t)}{c(V)} \exp\left(\tfrac{1}{2}\|S_t\|^2_{(V+V_t)^{-1}}\right) = \left(\frac{\det(V)}{\det(V+V_t)}\right)^{1/2} \exp\left(\tfrac{1}{2}\|S_t\|^2_{(V+V_t)^{-1}}\right)\, .
\end{aligned}$$

$c(P) = \sqrt{(2\pi)^d / \det(P)} = \int \exp(-\frac{1}{2}x^T P x)\, dx$

$M_t = \int_{\mathbb{R}^d} \exp(\langle \lambda, S_t \rangle - \frac{1}{2}\|\lambda\|^2_{V_t}) f(\lambda)\, d\lambda$

$= \int_{\mathbb{R}^d} \exp(-\frac{1}{2}\|\lambda - V_t^{-1}S_t\|^2_{V_t} + \frac{1}{2}\|S_t\|^2_{V_t^{-1}}) f(\lambda)\, d\lambda$

$\Rightarrow -\frac{1}{2}(\lambda - V_t^{-1}S_t)^T V_t (\lambda - V_t^{-1}S_t) + \frac{1}{2} S_t^T V_t^{-1} S_t$

$-\frac{1}{2}(\lambda^T - S_t^T \underline{V_t^{-1}}) V_t (\lambda - V_t^{-1}S_t) + \frac{1}{2} S_t^T V_t^{-1} S_t$

$-\frac{1}{2}(\lambda^T V_t - S_t^T)(\lambda - V_t^{-1}S_t) + \frac{1}{2} S_t^T V_t^{-1} S_t$

$-\frac{1}{2}(\lambda^T V_t \lambda - S_t^T\lambda - \lambda^T \underline{V_t V_t^{-1}} S_t + S_t \cancel{V_t^{-1}} S_t) + \frac{1}{2} \cancel{S_t^T} V_t^{-1} S_t$

$-\frac{1}{2}\|\lambda\|^2_{V_t} + \frac{1}{2} S_t^T \lambda + \frac{1}{2}\lambda^T S_t = \langle \lambda, S_t \rangle - \frac{1}{2}\|\lambda\|^2_{V_t}$

$= \frac{1}{c(V)} \exp(\frac{1}{2}\|S_t\|^2_{V_t^{-1}}) \int_{\mathbb{R}^d} \exp(-\frac{1}{2}\{\|\lambda - V_t^{-1}S_t\|^2_{V_t} + \|\lambda\|^2_V\})\, d\lambda$

V^{-1} 분산-공분산 행렬 → 대칭행렬

Theorem 1

Now, from $\mathbf{E}[M_\tau] \le 1$, we obtain

$$\Pr\left[\|S_\tau\|^2_{(V+V_\tau)^{-1}} > 2\log\left(\frac{\det(V+V_\tau)^{1/2}}{\delta\det(V)^{1/2}}\right)\right] = \Pr\left[\frac{\exp\left(\frac{1}{2}\ \|S_\tau\|^2_{(V+V_\tau)^{-1}}\right)}{\delta^{-1}\left(\det(V+V_\tau)\big/\det(V)\right)^{\frac{1}{2}}} > 1\right]$$

$$\le \mathbf{E}\left[\frac{\exp\left(\frac{1}{2}\ \|S_\tau\|^2_{(V+V_\tau)^{-1}}\right)}{\delta^{-1}\left(\det(V+V_\tau)\big/\det(V)\right)^{\frac{1}{2}}}\right]$$

$$= \mathbf{E}[M_\tau]\delta \le \delta,$$

$$= P(\delta M_\tau > 1) \le \frac{E(M_\tau)}{(\delta)^{-1}} \le \delta$$

thus finishing the proof. $\square$

4. Construction of Confidence Sets

- $\widehat{\theta}_t$ 는 규제 강도가 λ>0 정규화 파라미터를 가지는 L2 정규화 최소제곱법으로 구한 θ_* 의 예측치 (Ridge regression)

$$\widehat{\theta}_t = (\mathbf{X}_{1:t}^\top \mathbf{X}_{1:t} + \lambda I)^{-1} \mathbf{X}_{1:t}^\top \mathbf{Y}_{1:t}$$

$\mathbf{X}_{1:t}$ is the matrix whose rows are $X_1^\top, X_2^\top, \ldots, X_t^\top$ and $\mathbf{Y}_{1:t} = (Y_1, \ldots, Y_t)^\top$

- Theorem2는 θ_* 가 높은 확률로 $\widehat{\theta}_t$ 가 중심인 타원공간에 있음을 보인다.
- 새로운 신뢰 집합은 연산비용이 큰 행렬식 계산이 필요해 보이지만, matrix determinant lemma를 이용하면 속도를 높일 수 있다.(rank-one update)

Theorem 2

Theorem 2 (Confidence Ellipsoid). *Assume the same as in Theorem 1, let* $V = I\lambda$, $\lambda > 0$, *define* $Y_t = \langle X_t, \theta_* \rangle + \eta_t$ *and assume that* $\|\theta_*\|_2 \le S$. *Then, for any* $\delta > 0$, *with probability at least* $1 - \delta$, *for all* $t \ge 0$, θ_* *lies in the set*

$$C_t = \left\{ \theta \in \mathbb{R}^d \ : \ \left\| \widehat{\theta}_t - \theta \right\|_{\overline{V}_t} \le R \sqrt{2 \log \left(\frac{\det(\overline{V}_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right\} .$$

Furthermore, if for all $t \ge 1$, $\|X_t\|_2 \le L$ *then with probability at least* $1 - \delta$, *for all* $t \ge 0$, θ_* *lies in the set*

$$C'_t = \left\{ \theta \in \mathbb{R}^d \ : \ \left\| \widehat{\theta}_t - \theta \right\|_{\overline{V}_t} \le R \sqrt{d \log \left(\frac{1 + tL^2/\lambda}{\delta} \right)} + \lambda^{1/2} S \right\} .$$

5. Regret Analysis of the OFUL ALGORITHM

- CONFIDENCEBALL 알고리즘과 유사하지만 350배 연산을 덜하고 더 좋은 regret을 보여준다.

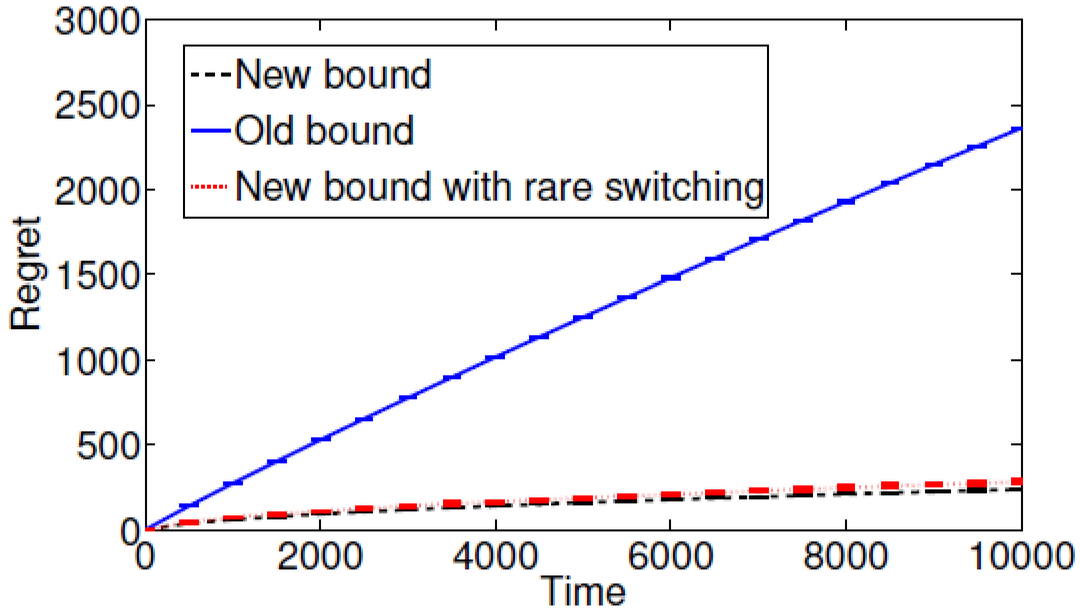
Theorem 3 (The regret of the OFUL algorithm)**.** *Assume that for all t and all $x \in D_t$, $\langle x, \theta_* \rangle \in [-1, 1]$. Then, with probability at least $1 - \delta$, the regret of the* OFUL *algorithm satisfies*

$$\forall n \geq 0, \quad R_n \leq 4\sqrt{nd \log(\lambda + nL/d)} \left(\lambda^{1/2} S + R\sqrt{2 \log(1/\delta) + d \log(1 + nL/(\lambda d))} \right) .$$

Theorem 3

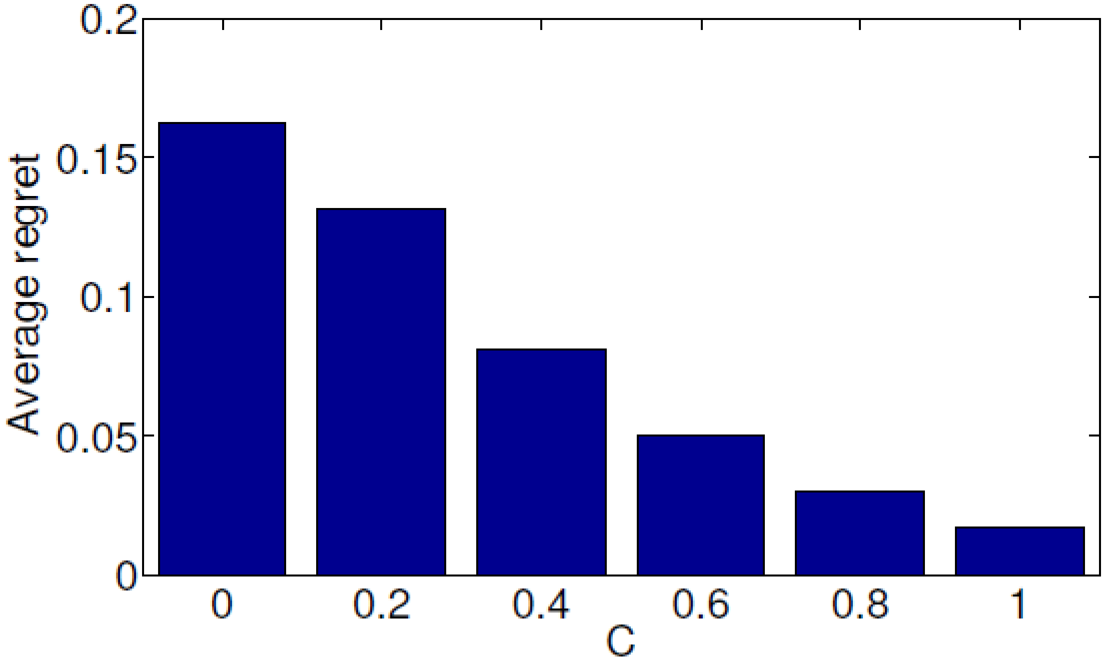
5.1 Saving Computation

- det(Vt)가 C+1만큼 증가할 때마다 $\tilde{\theta}_t$ 를 재계산한다.



- CONFIDENCEBALL 에 비해 regret이 눈에 띄게 나아졌다.
- noise는 평균이 0이고 표준편차가 0.1인 정규분포다.
- 신뢰 집합 밖에 있을 확률은 0.0001이다.

5.1 Saving Computation



- C가 커질수록 알고리즘은 행동을 덜 자주 변경한다. 따라서 오랜 기간 작업을 지속할 수 있다.
- 정해진 연산 budget 내에서 시간당 평균 regret을 낮출 수 있다.

Theorem 4

Theorem 4. *Under the same assumptions as in Theorem 3, with probability at least* $1-\delta$, *for all* $n \geq 0$, *the regret of the* RARELY SWITCHING OFUL ALGORITHM *satisfies*

$$R_n \leq 4\sqrt{(1+C)nd\log\left(\lambda+\frac{nL}{d}\right)}\left\{\sqrt{\lambda}S+R\sqrt{d\log\left(1+\frac{nL}{\lambda d}\right)+2\log\frac{1}{\delta}}\right\}+4\sqrt{d\log\frac{n}{d}}\,.$$

6. Multi-Armed Bandit Problem

- μ_i -> action i=1,2,….,d 의 expected reward
- μ_* -> 최적 arm의 expected reward
- $\Delta_i = \mu_* - \mu_i,\ i = 1, 2, \ldots, d,$
- $\mu_{I_t} + \eta_t$ -> t라운드에 I_t행동을 했을 때 얻는 reward
- $N_{i,t}$ -> t시점까지 action i를 시행한 횟수
- $\overline{X}_{i,t}$ -> t시점까지 action i를 했을 때 얻는 reward
- μ_i 의 신뢰 구간을 $\overline{X}_{i,t}$ 에 근거하여 구할 수 있다.

6. Multi-Armed Bandit Problem

Lemma 6 (Confidence Intervals). *Assuming that the noise η_t is conditionally 1-sub-Gaussian. With probability at least $1-\delta$,*

$$\forall i \in \{1,2,\ldots,d\},\ \forall t \geq 0 \qquad |\overline{X}_{i,t} - \mu_i| \leq c_{i,t}\ ,$$

where

$$c_{i,t} = \sqrt{\frac{(1+N_{i,t})}{N_{i,t}^2}\left(1 + 2\log\left(\frac{d(1+N_{i,t})^{1/2}}{\delta}\right)\right)}\ . \tag{3}$$

- 이 신뢰 구간을 이용하여, UCB알고리즘을 수정하고 action selection rule을 변경했다.
- UCB(δ), $I_t = \operatorname*{argmax}_i \overline{X}_{i,t} + c_{i,t}$
- UCB(δ)와 UCB의 차이점은 신뢰 구간이 n과 t에 의존하지 않는다는 것.

(Appendix G)

- 새로운 bound는 union bound를 피할 수 있는 new self-normalized tail inequality로 더 타이트해졌다(좁아졌다).

7. Conclusions

- 어떻게 새로운 vector-valued martingale에 대한 tail inequality가 다양한 stochastic bandit problem의 알고리즘의 이론적인 분석과 실증적인 성능 모두를 향상시킬 수 있는지 보였다.
- Auer's UCB algorithm의 간단한 수정을 통해 높은 확률로 constant regret을 얻는 것을 보였다.
- Auer, Dani, Rusmevichientong and Tsitsiklis, Li가 연구한 linear stochastic bandit problem에 대한 알고리즘의 분석을 수정하고 개선했다.
- logarithmic factor에 의한 regret bound의 개선을 보여준다
- 성능의 희생없이 많은 양의 연산을 줄일 수 있다.
- 새로운 부등식은 stopped martingale에 사용되며, bound를 균일하게 할 수 있다. union bound를 사용하는 deviation bound들의 개선하는데 사용할 수 있다.
- 현대의 많은 머신러닝 모델들이 high-probability bound에 의존하기 때문에 새로운 부등식이 많이 쓰일 수 있다.