Linear Bandits

Sham M. Kakade (and Wen Sun)

Hi!

Outline

1. Recap
2. Linear Bandits
   - Setting
   - LinUCB
   - An Optimal Regret Bound
3. Analysis
   - Regret Analysis
   - Confidence Analysis

Generalization in RL

- (distribution free) Agnostic learning is not possible in RL:
  we showed that to get $O(\log|\Pi|)$ sample complexity we need either:
  - $\mathrm{poly}(|\mathcal{S}|)$ samples OR
  - $\cancel{\mathrm{poly}}(H)$ samples.
    $\exp(H)$

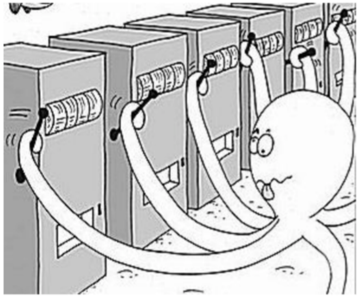
  in order to learn the best policy in some policy class.
- upshot: we need stronger assumptions for RL analysis.

Multi-Armed-Bandits: High-level picture

Setting

- Set of alternatives (arms)
- Each arm has a reward distribution

- Learner adaptively selects arms
- Challenge: Distributions not known



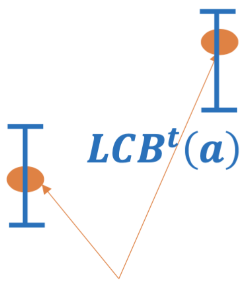
Images from:
https://towardsdatascience.com/beyond-a-b-testing-multi-armed-bandit-experiments-1493f709f804
https://www.aqusagtechnologies.com/wp-content/uploads/2017/04/Online-advertising.jpeg

Upper Confidence Bound (UCB)

Pick arm with highest Upper Confidence Bound

$$UCB^t(a) = \tilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$$

$$LCB^t(a) = \tilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$$



By Hoeffding and union bound, with probability $\geq 1-\delta$, it holds $\forall a \in [k], t \in [T]$:

$$\mu(a) \in [LCB^t(a), UCB^t(a)]$$

Claim : In the event that all confidence intervals confidence intervals hold, the regret is at most $\sum_t (UCB^t(a^t) - LCB^t(a^t)) + \delta \cdot T$

Proof: $Reg^t = \mu(a^\star) - \mu(a^t)$

$$\leq UCB^t(a^\star) - LCB^t(a^t)$$
$$\leq UCB^t(a^t) - LCB^t(a^t)$$

Upper Confidence Bound (UCB)

$$UCB^t(a) = \tilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}} \qquad LCB^t(a) = \tilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$$

Claim : In the event that all confidence intervals confidence intervals hold, the regret is at most $\sum_t (UCB^t(a^t) - LCB^t(a^t)) + +\delta \cdot T$

Regret bound by confidence sum

$$\sum_t (UCB^t(a^t) - LCB^t(a^t)) \leq 2 \cdot \sum_t \sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2n^t(a^t)}} = \sum_a \sum_{j=1}^{N(a)} \sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2 \cdot j}}$$

$$\leq \sum_a \sum_{j=1}^{\frac{T}{k}} \sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2 \cdot j}} \leq k \cdot \sqrt{\log\left(\frac{2kT}{\delta}\right) \cdot \frac{T}{k}} = O\left(\sqrt{T \cdot k \cdot \log\left(\frac{kT}{\delta}\right)}\right)$$

regret $\leq O(\sqrt{T \cdot \# \text{arms}})$

Outline

1. Recap
2. Linear Bandits
   - Setting
   - LinUCB
   - An Optimal Regret Bound
3. Analysis
   - Regret Analysis
   - Confidence Analysis

Handling Large Actions Spaces

Lin Bandit Model

Abe & Long '99

- On each round, we must choose a decision $x_t \in D \subset R^d$.

Handling Large Actions Spaces

- On each round, we must choose a decision $x_t \in D \subset R^d$.
- Obtain a reward $r_t \in [-1, 1]$, where

$$\mathbb{E}[r_t | x_t = x] = \mu^\star \cdot x \in [-1, 1],$$

Handling Large Actions Spaces

- On each round, we must choose a decision $x_t \in D \subset R^d$.
- Obtain a reward $r_t \in [-1, 1]$, where

$$\mathbb{E}[r_t | x_t = x] = \mu^\star \cdot x \in [-1, 1],$$

  - so the the conditional expectation of r_t is linear)
  - Also, we have the *noise sequence*,

$$\eta_t = r_t - \mu^\star \cdot x_t$$

  is i.i.d noise.

observe

$r_t = \mu^* \cdot X + \eta$

unknown

Our Objective

$$\sum_{t=0}^{T-1} r_t \approx \sum_{t=0}^{T-1} \mu^\star \cdot x_t$$

If $x_0, \ldots x_{T-1}$ are our decisions, then our cumulative regret is

$$R_T = T(\mu^\star \cdot x^\star) - \sum_{t=0}^{T-1} \mu^\star \cdot x_t$$

online Lin. Programming

where $x^\star \in D$ is an optimal decision for $\mu^\star$, i.e.

$$x^\star \in \mathrm{argmax}_{x \in D}\, \mu^\star \cdot x$$

if D was a polytope.

Outline

The "Confidence Ball"

After t rounds, define our uncertainty region $\textsc{Ball}_t$: with center, $\widehat{\mu}_t$, and shape, Σ_t, using the λ-regularized least squares solution:

seen

$x_1 \dots x_{t-1}$

$r_1 \dots r_{t-1}$

$$\widehat{\mu}_t = \arg\min_{\mu} \sum_{\tau=0}^{t-1} \|\mu \cdot x_t - r_t\|_2^2 + \lambda \|\mu\|_2^2$$

$$= \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_\tau x_\tau ,$$

$$\Sigma_t = \lambda I + \sum_{\tau=0}^{t-1} x_t x_t^\top , \text{ with } \Sigma_0 = \lambda I .$$

The "Confidence Ball"

After t rounds, define our uncertainty region $\textsc{Ball}_t$: with center, $\widehat{\mu}_t$, and shape, Σ_t, using the λ-regularized least squares solution:

$$\widehat{\mu}_t = \arg\min_{\mu} \sum_{\tau=0}^{t-1} \|\mu \cdot x_t - r_t\|_2^2 + \lambda \|\mu\|_2^2$$

$$= \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_\tau x_\tau ,$$

$$\Sigma_t = \lambda I + \sum_{\tau=0}^{t-1} x_t x_t^\top , \text{ with } \Sigma_0 = \lambda I.$$

Define the uncertainty region: $\{\mu \mid (\hat{\mu}_t - \mu) \Sigma_t (\hat{\mu}_t - \mu) \le \beta_t\}$

$$\textsc{Ball}_t = \left\{ (\widehat{\mu}_t - \mu^\star)^\top \Sigma_t^{-1} (\widehat{\mu}_t - \mu^\star) \le \beta_t \right\},$$

where β_t is a parameter of the algorithm.

LinUCB (the algo)

1. Input: λ, β_t
2. For $t = 0, 1, \ldots$
   1. Execute
      $$x_t = \mathrm{argmax}_{x \in D} \max_{w \in \textsc{Ball}_t} w \cdot x$$
      and observe the reward r_t.
   2. Update $\textsc{Ball}_{t+1}$.

using x_t, r_t

LinUCB Regret Bound

$|D| \approx (\frac{1}{\epsilon})^d$

Sublinear regret: $R_T \leq O^\star(d\sqrt{T})$
poly dependence on d , no dependence on the cardinality $|D|$.

LinUCB Regret Bound

Sublinear regret: $R_T \leq O^\star(d\sqrt{T})$ wlog ∃ scaling $W \approx O(\sqrt{d})$

poly dependence on d , no dependence on the cardinality $|D|$. $\|x\| \leq 1$

Theorem

Suppose: bounded noise $|\eta_t| \leq \sigma$, that $\|\mu^\star\| \leq W$, and that $\|x\| \leq B$ for all $x \in D$. Set $\lambda = \sigma^2/W^2$ and

$$\beta_t := \sigma^2\left(2 + 4d\log\left(1 + \frac{\overset{t}{\cancel{T}}B^2W^2}{d}\right) + 8\log(4/\delta)\right).$$

With probability greater than $1-\delta$, that for all $t \geq 0$,

$$R_T \leq c\sigma\sqrt{T}\left(d\log\left(1 + \frac{TB^2W^2}{d\sigma^2}\right) + \log(4/\delta)\right)$$

where c is an absolute constant.

Tight in d & T

(due to Dani, K., and Hayes '09)

av reg. $d/\sqrt{T}$

Outline

Confidence

In establishing the upper bounds there are two main propositions from which the upper bounds follow. The first is in showing that the confidence region is valid.

Proposition

(Confidence) Let $\delta > 0$. We have that

$$\Pr(\forall t,\ \mu^{\star} \in \textsc{Ball}_t) \geq 1 - \delta.$$

Sum of Squares Regret Bound

Assuming the confidence event holds, the following controls on the growth of the regret.

Proposition

(Sum of Squares Regret Bound) Define: instantaneous regret

$$\text{regret}_t = \mu^\star \cdot x^* - \mu^\star \cdot x_t$$

Suppose $\|x\| \leq B$ *for* $x \in D$*. Suppose* β_t *is set as in Theorem 1. Suppose* $\mu^\star \in \textsc{Ball}_t$ *for all* t*, then* $O(\cdot \log T)$

$$\sum_{t=0}^{T-1} \text{regret}_t^2 \leq 4\beta_T d \log\left(1 + \frac{TB^2}{d\lambda}\right)$$

Completing the Proof

$\beta_T \tilde{\approx} O(d \log T)$

Proof:[Proof of Theorem 1] With the two previous Propositions, along with the Cauchy-Schwarz inequality, we have, with probability at least $1 - \delta$,

$$R_T = \sum_{t=0}^{T-1} \mathrm{regret}_t \le \left(T \sum_{t=0}^{T-1} \mathrm{regret}_t^2 \right)^{1/2} \le \sqrt{4 T \beta_T d \log\left(1 + \frac{TB^2}{d\lambda}\right)}.$$

The remainder of the proof follows from using our chosen value of β_T and algebraic manipulations. ■

Outline

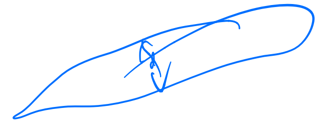
"Width" of Confidence Ball

Lemma

Let $x \in D$. If $\mu \in \textsc{Ball}_t$ and $x \in D$. Then

$$|(\mu - \widehat{\mu}_t)^\top x| \leq \sqrt{\beta_t x^\top \Sigma_t^{-1} x}$$

“Width” of Confidence Ball



Lemma

Let $x \in D$. If $\mu \in \text{BALL}_t$ and $x \in D$. Then

$$|(\mu - \widehat{\mu}_t)^\top x| \le \sqrt{\beta_t x^\top \Sigma_t^{-1} x}$$

Proof: By Cauchy-Schwarz, we have:

$$|(\mu - \widehat{\mu}_t)^\top x| = |(\mu - \widehat{\mu}_t)^\top \Sigma_t^{1/2} \Sigma_t^{-1/2} x| = |(\Sigma_t^{1/2}(\mu - \widehat{\mu}_t))^\top \Sigma_t^{-1/2} x|$$

$$\le \|\Sigma_t^{1/2}(\mu - \widehat{\mu}_t)\| \|\Sigma_t^{-1/2} x\| = \|\Sigma_t^{1/2}(\mu - \widehat{\mu}_t)\| \sqrt{x^\top \Sigma_t^{-1} x} \le \sqrt{\beta_t x^\top \Sigma_t^{-1} x}$$

where the last inequality holds since $\mu \in \text{BALL}_t$. ■

Instantaneous Regret Lemma

Define

$$w_t := \sqrt{x_t^\top \Sigma_t^{-1} x_t}$$

which is the “normalized width” at time t in the direction of our decision.

Instantaneous Regret Lemma

Define

$$w_t := \sqrt{x_t^\top \Sigma_t^{-1} x_t}$$

which is the "normalized width" at time t in the direction of our decision.

Lemma

Fix $t \le T$. If $\mu^\star \in \textsc{Ball}_t$, then

$$\text{regret}_t \le 2 \min\left(\sqrt{\beta_t} w_t, 1\right) \le 2\sqrt{\beta_T} \min\left(w_t, 1\right)$$

β_t is increasing

Instantaneous Regret Lemma

Define

$\Sigma_{t+1} \leftarrow \Sigma_t + x_t x_t^\top$

$$w_t := \sqrt{x_t^\top \Sigma_t^{-1} x_t}$$

which is the "normalized width" at time t in the direction of our decision.

Lemma

Fix $t \leq T$. If $\mu^\star \in \textsc{Ball}_t$, then

$E[r_t | x_t] \in [-1, 1]$

$$\mathrm{regret}_t \leq 2 \min\left(\sqrt{\beta_t} w_t, 1\right) \leq 2\sqrt{\beta_T} \min\left(w_t, 1\right)$$

Proof: Let $\widetilde{\mu} \in \textsc{Ball}_t$ denote the vector which minimizes the dot product $\widetilde{\mu}^\top x_t$. By choice of x_t, we have

$$\widetilde{\mu}^\top x_t = \max_{\mu \in \textsc{Ball}_t} \max_{x \in D} \mu^\top x \geq (\mu^\star)^\top x^*,$$

where the inequality used the hypothesis $\mu^\star \in \textsc{Ball}_t$. Hence,

$$\begin{aligned} \mathrm{regret}_t &= (\mu^\star)^\top x^* - (\mu^\star)^\top x_t \leq (\widetilde{\mu} - \mu^\star)^\top x_t \\ &= (\widetilde{\mu} - \widehat{\mu}_t)^\top x_t + (\widehat{\mu}_t - \mu^\star)^\top x_t \leq 2\sqrt{\beta_t} w_t \end{aligned}$$

Geometric Argument: Part 1

The next two lemmas give us 'geometric' potential function argument, where can bound the sum of widths independently of the choices made by the algorithm.

Geometric Argument: Part 1

The next two lemmas give us 'geometric' potential function argument, where can bound the sum of widths independently of the choices made by the algorithm.

Lemma

We have:

$$\det \Sigma_T = \det \Sigma_0 \prod_{t=0}^{T-1} (1 + w_t^2).$$

Geometric Argument: Part 1

The next two lemmas give us 'geometric' potential function argument, where can bound the sum of widths independently of the choices made by the algorithm.

$w_t = \sqrt{x_t \Sigma_t^{-1} x_t}$

Lemma

We have:

volume update equality

$$\det \Sigma_T = \det \Sigma_0 \prod_{t=0}^{T-1} (1 + w_t^2).$$

Proof: By the definition of Σ_{t+1}, we have

$$\det \Sigma_{t+1} = \det(\Sigma_t + x_t x_t^\top) = \det(\Sigma_t^{1/2}(I + \Sigma_t^{-1/2} x_t x_t^\top \Sigma_t^{-1/2})\Sigma_t^{1/2})$$
$$= \det(\Sigma_t)\det(I + \Sigma_t^{-1/2} x_t (\Sigma_t^{-1/2} x_t)^\top) = \det(\Sigma_t)\det(I + v_t v_t^\top),$$

where $v_t := \Sigma_t^{-1/2} x_t$. Now observe that $v_t^\top v_t = w_t^2$ and ... ■

Geometric Argument: Part 2

Lemma

For any sequence $x_0, \ldots x_{T-1}$ such that, for $t < T$, $\|x_t\|_2 \leq B$, we have:

$\Sigma_0 \succeq \lambda I$

$$\log\left(\det \Sigma_{T-1} / \det \Sigma_0\right) = \log\det\left(I + \frac{1}{\lambda}\sum_{t=0}^{T-1} x_t x_t^\top\right) \leq d \log\left(1 + \frac{TB^2}{d\lambda}\right).$$

Geometric Argument: Part 2

Lemma

For any sequence $x_0, \ldots x_{T-1}$ such that, for $t < T$, $\|x_t\|_2 \le B$, we have:

$$\log\Big(\det \Sigma_{T-1} / \det \Sigma_0\Big) = \log\det\left(I + \frac{1}{\lambda}\sum_{t=0}^{T-1} x_t x_t^\top\right) \le d \log\left(1 + \frac{TB^2}{d\lambda}\right).$$

Proof: Denote the eigenvalues of $\sum_{t=0}^{T-1} x_t x_t^\top$ as $\sigma_1, \ldots \sigma_d$, and note:

$$\sum_{i=1}^{d} \sigma_i = \mathrm{Trace}\Big(\sum_{t=0}^{T-1} x_t x_t^\top\Big) = \sum_{t=0}^{T-1} \|x_t\|^2 \le TB^2.$$

Using the AM-GM inequality,

$$\log\det\left(I + \frac{1}{\lambda}\sum_{t=0}^{T-1} x_t x_t^\top\right) = \log\left(\prod_{i=1}^{d}(1 + \sigma_i/\lambda)\right)$$

$$= d\log\left(\prod_{i=1}^{d}\Big(1 + \sigma_i/\lambda\Big)\right)^{1/d} \le d\log\Big(\frac{1}{d}\sum_{i=1}^{d}(1 + \sigma_i/\lambda)\Big) \le d\log\left(1 + \frac{TB^2}{d\lambda}\right)$$

Proving "sum of squares regret" Proposition

Proof:[Proof of Proposition 3] Assume $\mu^\star \in \textsc{Ball}_t$ for all t. We have:

$$
\begin{aligned}
\sum_{t=0}^{T-1} \text{regret}_t^2 &\le \sum_{t=0}^{T-1} \overset{2}{\cancel{4}}\beta_t \min(w_t^2, 1) \le \overset{2}{\cancel{4}}\beta_T \sum_{t=0}^{T-1} \min(w_t^2, 1) \\
&\le 4\beta_T \sum_{t=0}^{T-1} \ln(1 + w_\tau^2) \le 4\beta_T \log\left(\det \Sigma_{T-1} / \det \Sigma_0\right) \\
&= 4\beta_T d \log\left(1 + \frac{TB^2}{d\lambda}\right)
\end{aligned}
$$

where the first inequality follow from by Lemma 5; the second from that β_t is an increasing function of t; the third uses that for $0 \le y \le 1$, $\ln(1 + y) \ge y/2$; the final two inequalities follow by Lemmas 6 and 7. ∎

Outline

1. Recap
2. Linear Bandits
   - Setting
   - LinUCB
   - An Optimal Regret Bound
3. Analysis
   - Regret Analysis
   - **Confidence Analysis**

$\Pr(\forall t, \mu^* \in Ball_t)$

$\leq \delta$

Confidence [Proof of Proposition 2]

Proof: Since $r_\tau = x_\tau \cdot \mu^\star + \eta_\tau$, we have:

$$\widehat{\mu}_t - \mu^\star = \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_\tau x_\tau - \mu^\star = \Sigma_t^{-1} \sum_{\tau=0}^{t-1} x_\tau (x_\tau \cdot \mu^\star + \eta_\tau) - \mu^\star$$

$$= \Sigma_t^{-1} \left(\sum_{\tau=0}^{t-1} x_\tau (x_\tau)^\top \right) \mu^\star - \mu^\star + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau$$

$+ \lambda I - \lambda I$

$$= \lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau$$

$\Sigma_t = \sum x_\tau x_\tau^\top + \lambda I$

Confidence [Proof of Proposition 2]

Proof: Since $r_\tau = x_\tau \cdot \mu^\star + \eta_\tau$, we have:

$$\widehat{\mu}_t - \mu^\star = \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_\tau x_\tau - \mu^\star = \Sigma_t^{-1} \sum_{\tau=0}^{t-1} x_\tau (x_\tau \cdot \mu^\star + \eta_\tau) - \mu^\star$$

$$= \Sigma_t^{-1} \left(\sum_{\tau=0}^{t-1} x_\tau (x_\tau)^\top \right) \mu^\star - \mu^\star + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau$$

$$= \lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau$$

By the triangle inequality,

$\Sigma_t^{-1/2}$

$$(\widehat{\mu}_t - \mu^\star)^\top \Sigma_t^{-1} (\widehat{\mu}_t - \mu^\star) \le \left\| \lambda (\Sigma_t)^{-1/2} \mu^\star \right\| + \left\| \cancel{\Sigma_t^{-1}} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \right\|$$

$$\le \sqrt{\lambda} \|\mu^\star\| \qquad + \qquad ??.$$

How can we bound “??” To be continued... ■

Self-Normalizing Sum

$\xi_i \longleftrightarrow \eta_i$

$V_t \longleftrightarrow \Sigma_t$

Lemma (Self-Normalized Bound for Vector-Valued Martingales)

(Abassi et. al '11) Suppose $\{\varepsilon_i\}_{i=1}^{\infty}$ *are mean zero random variables (can be generalized to martingales), and* ε_i *is bounded by* σ. *Let* $\{X_i\}_{i=1}^{\infty}$ *be a stochastic process. Define* $V_t = V_0 + \sum_{i=1}^{t} X_i X_i^\top$. *With probability at least* $1 - \delta$, *we have for all* $t \geq 1$:

$\det(\Sigma_t) / \det(\Sigma_0)$

$$\left\| \sum_{i=1}^{t} X_i \varepsilon_i \right\|_{V_t^{-1}}^2 \leq 2\sigma^2 \log \left(\frac{\det(V_t)^{1/2} \det(V_0)^{-1/2}}{\delta^2} \right).$$

Continued... [Proof of Proposition 2]

Proof:

$$\sqrt{(\widehat{\mu}_t - \mu^\star)^\top \Sigma_t^{-1} (\widehat{\mu}_t - \mu^\star)} \leq \left\| \lambda(\Sigma_t)^{-1/2} \mu^\star \right\| + \left\| \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \right\|$$

$$\leq \sqrt{\lambda} \|\mu^\star\| + \sqrt{2\sigma^2 \log\left(\det(\Sigma_t) \det(\Sigma^0)^{-1} / \delta_t\right)}.$$

We seek to lower bound $\Pr(\forall t,\ \mu^\star \in \textsc{Ball}_t)$. Assign failure probability $\delta_t = (3/\pi^2)/t^2$ for the t-th event, which gives us:

$$1 - \Pr(\forall t,\ \mu^\star \in \textsc{Ball}_t) = \Pr(\exists t,\ \mu^\star \notin \textsc{Ball}_t) \leq \sum_{t=1}^{\infty} \Pr(\mu^\star \notin \textsc{Ball}_t)$$

$$< \sum_{t=1}^{\infty} (1/t^2)(3/\pi^2) = 1/2.$$

This along with Lemma 7 completes the proof. ∎