Bandit Algorithms (part 1)

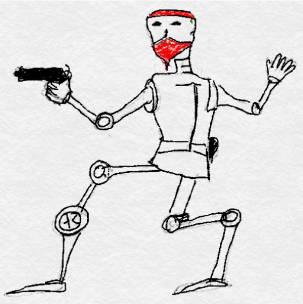
Tor Lattimore

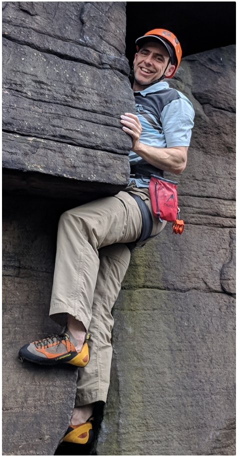
‘Bandit Algorithms’ book

Joint work with Csaba

Free online at `http://banditalgs.com`

Covers all topics in slides and more





Overview

Today

- What are bandits
- Applications
- Optimism in the face of uncertainty
- Scaling up
- Linear bandits and structure

Next

- Adversarial bandits
- Online convex optimization
- Mirror descent
- Bandits, combinatorial bandits, shortest path problems, adversarial linear bandits

Bandit problems

- Baby reinforcement learning
- Acting in the face of uncertainty
- No planning

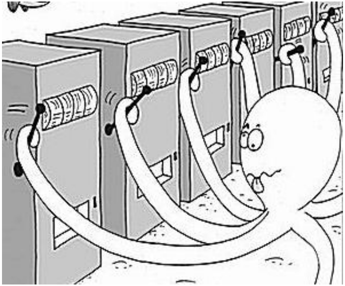
Bandits

Finite action set $\mathcal{A} = \{1, 2, \ldots, k\}$

For each $a \in \mathcal{A}$ there is an **unknown** distribution P_a

Learner chooses $A_t \in \mathcal{A}$ and observes **reward** $R_t \sim P_{A_t}$

Learner wants to maximise $\sum_{t=1}^{n} R_t$



Why care?

- A simplified view of exploration/exploitation
- Applications
- Fun math

Applications

- Clinical trials
- A/B testing
- Ad placement
- Recommender systems
- Network routing
- Game tree search

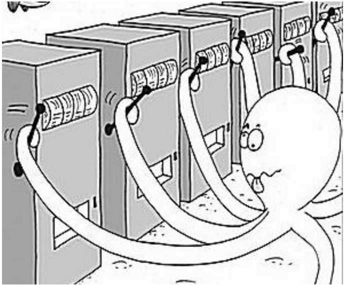
Bandits

Finite action set $\mathcal{A} = \{1, 2, \ldots, k\}$

For each $a \in \mathcal{A}$ there is an **unknown** distribution P_a

Learner chooses $A_t \in \mathcal{A}$ and observes **reward** $R_t \sim P_{A_t}$

Learner wants to maximise $\sum_{t=1}^{n} R_t$



The learning objective

Let μ_a be the mean of P_a and $\mu^* = \max_{a \in \mathcal{A}} \mu_a$

The **optimal action** is $a^* = \mathrm{argmax}_a \, \mu_a$

Our task is to minimise the **regret**

$$\mathfrak{R}_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right]$$

The price paid by the learner for not knowing μ

Assumptions matter

Mean reward for each arm are **unknown**

Necessary to make some assumptions

Examples:

- Bernoulli
- Gaussian with unknown mean and unit variance
- Gaussian with unknown mean and unknown variance
- 1-subgaussian
- Bounded in $[0, 1]$ with unknown variance
- Supported on $(-\infty, b]$
- Unknown mean and variance less than known σ^2
- Kurtosis less than κ
- Many more

Strong assumptions lead to better algorithms **(if you're right)**

Algorithmic idea

Estimate the mean of each arm

Only play arms that are **statistically plausibly** optimal

What is this ‘statistically plausible’ and which arm to play?

We need our assumptions. For the next little while: **Gaussian with unit variance**

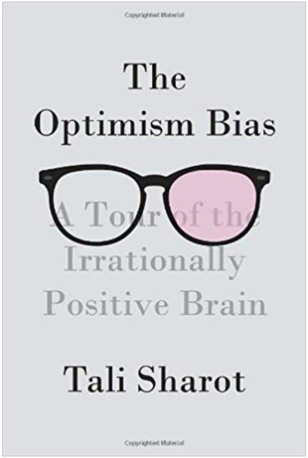
Optimism

People are naturally optimistic

Psychological benefits and...

Encourages exploration

(some downsides too)



Optimism principle

‘You should act as if you are in the **nicest plausible** world possible’

Optimism principle

‘You should act as if you are in the **nicest plausible** world possible’

Guarantees either (a) **optimality** or (b) **exploration**

Concentration for Gaussian sums

Let $X_1, \ldots, X_T$ be a sequence of independent Gaussian random variables with mean μ and variance 1 and

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} X_t$$

Then for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{T}}\right) \leq \delta$$

$$\mathbb{P}\left(\hat{\mu} \leq \mu - \sqrt{\frac{2\log(1/\delta)}{T}}\right) \leq \delta$$

‘Nicest’ In bandits, we want the mean to be large

‘Plausible’ The mean cannot be *much* larger than the empirical mean

'Nicest' In bandits, we want the mean to be large

'Plausible' The mean cannot be *much* larger than the empirical mean

Upper Confidence Bound Algorithm

Choose each arm once and then

$$A_t = \text{argmax}_a \, \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}}$$

$\hat{\mu}_a(t) =$ empirical mean of arm a after round t

$T_a(t) =$ number of plays of arm a after round t

$\delta =$ confidence level

Regret analysis

Step 1 Decompose the regret over the arms

Step 2 On a ‘good’ event prove that suboptimal arms are not played too often

Step 3 Show the ‘good’ event occurs with high probability

Regret decomposition

$$\mathfrak{R}_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right]$$

$\Delta_a = \mu^* - \mu_a$

$T_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$

Regret decomposition

$\Delta_a = \mu^* - \mu_a$

$T_a(t) = \sum_{s=1}^t \mathbb{1}(A_s = a)$

$$\begin{aligned}\mathfrak{R}_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^n R_t\right] \\ &= \mathbb{E}\left[\sum_{t=1}^n (\mu^* - R_t)\right]\end{aligned}$$

Regret decomposition

$$\Delta_a = \mu^* - \mu_a$$

$$T_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$$

$$\begin{aligned}
\mathfrak{R}_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} (\mu^* - R_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}\right]
\end{aligned}$$

Regret decomposition

$$\Delta_a = \mu^* - \mu_a$$

$$T_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$$

$$\begin{aligned}
\mathfrak{R}_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} (\mu^* - R_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{1}(A_t = a)\Delta_a\right]
\end{aligned}$$

Regret decomposition

$$\Delta_a = \mu^* - \mu_a$$

$$T_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$$

$$\begin{aligned}
\mathfrak{R}_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}(\mu^* - R_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} \mathbb{1}(A_t = a)\Delta_a\right] \\
&= \sum_{a\in\mathcal{A}} \Delta_a \mathbb{E}[T_a(n)]
\end{aligned}$$

Assume for all t that

$$\mu_a + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \hat{\mu}_a(t-1)$$

$$\hat{\mu}_{a^*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_{a^*}(t-1)}} \geq \mu^*$$

Assume for all t that

$$\mu_a + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \hat{\mu}_a(t-1)$$

$$\hat{\mu}_{a^*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_{a^*}(t-1)}} \geq \mu^*$$

Now suppose that $A_t = a$ in round t

$$\mu_a + 2\sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}}$$

Assume for all t that $$\mu_a + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \hat{\mu}_a(t-1)$$

$$\hat{\mu}_{a^*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_{a^*}(t-1)}} \geq \mu^*$$

Now suppose that $A_t = a$ in round t

$$\mu_a + 2\sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}}$$

$$\geq \hat{\mu}_{a^*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_{a^*}(t-1)}} \geq \mu_{a^*} = \mu_a + \Delta_a$$

Assume for all t that

$$\mu_a + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \hat{\mu}_a(t-1)$$

$$\hat{\mu}_{a^*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_{a^*}(t-1)}} \geq \mu^*$$

Now suppose that $A_t = a$ in round t

$$\mu_a + 2\sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}}$$

$$\geq \hat{\mu}_{a^*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_{a^*}(t-1)}} \geq \mu_{a^*} = \mu_a + \Delta_a$$

Hence

$$T_a(t-1) \leq \frac{8\log(1/\delta)}{\Delta_a^2} \implies T_a(n) \leq 1 + \frac{8\log(1/\delta)}{\Delta_a^2}$$

Let $\hat{\mu}_{a,s}$ be the empirical mean of arm a after s plays

The concentration theorem shows that

$$\mathbb{P}\left(\hat{\mu}_{a,s} \geq \mu_a + \sqrt{\frac{2\log(1/\delta)}{s}}\right) \leq \delta$$

Combining with a union bound,

$$\mathbb{P}\left(\text{exists } s \leq n : \hat{\mu}_{a,s} \geq \mu_a + \sqrt{\frac{2\log(1/\delta)}{s}}\right) \leq n\delta$$

$$\mathbb{P}\left(\cup_i B_i\right) \leq \sum_i \mathbb{P}\left(B_i\right)$$

Putting it together

$$
\begin{aligned}
\mathfrak{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\
&\leq \sum_{a \in \mathcal{A}: \Delta_a > 0} \Delta_a \left(2\delta n^2 + 1 + \frac{8 \log(1/\delta)}{\Delta_a^2} \right) \\
&\leq \sum_{a \in \mathcal{A}: \Delta_a > 0} 3\Delta_a + \frac{16 \log(n)}{\Delta_a}
\end{aligned}
$$

Choose $\delta = 1/n^2$

Sanity checking our results

We have proven the regret of UCB is at most

$$\mathfrak{R}_n \leq \sum_{a \in \mathcal{A}: \Delta_a > 0} 3\Delta_a + \frac{16 \log(n)}{\Delta_a}$$

Useless when Δ is very small

Problem independent bound

$$\mathfrak{R}_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)]$$

Problem independent bound

$$
\begin{aligned}
\mathfrak{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\
&= \sum_{a \in \mathcal{A} : \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in \mathcal{A} : \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)]
\end{aligned}
$$

Problem independent bound

$$
\begin{aligned}
\mathfrak{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\
&= \sum_{a \in \mathcal{A} : \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in \mathcal{A} : \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)] \\
&\leq n\Delta + \sum_{a \in \mathcal{A} : \Delta_a > \Delta} 3\Delta_a + \frac{16 \log(n)}{\Delta_a}
\end{aligned}
$$

Problem independent bound

$$
\begin{aligned}
\mathfrak{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\
&= \sum_{a \in \mathcal{A} : \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in \mathcal{A} : \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)] \\
&\leq n\Delta + \sum_{a \in \mathcal{A} : \Delta_a > \Delta} 3\Delta_a + \frac{16 \log(n)}{\Delta_a} \\
&= O(\sqrt{nk \log(n)})
\end{aligned}
$$

Refinements

- Anytime algorithm: $\hat{\mu}_i(t-1) + \sqrt{\dfrac{2\,\overline{\log}(t)}{T_i(t-1)}}$

  where $\overline{\log}(t) = \log(1 + t\log^2(t))$

- Optimal constants: $\limsup_{n\to\infty} \dfrac{\mathfrak{R}_n}{\log(n)} = \sum_{i:\Delta_i>0} \dfrac{2}{\Delta_i}$

- Minimax optimality: $\mathfrak{R}_n = O(\sqrt{kn})$

  (for a different algorithm)

- Lower bounds

Limitations

- Model is not practical when k is very large
- Lot's of bandit problems exhibit structure
  - Many ad's look similar
  - Routes in a network share paths
- Need to introduce some structure

Contextual linear bandits

- Action set is $\mathcal{A}_t \subset \mathbb{R}^d$
- Choose action $A_t \in \mathcal{A}_t$
- Reward is $X_t = \langle A_t, \theta \rangle + \eta_t$
- $\theta \in \mathbb{R}^d$ is unknown
- η_t is the noise
- Lots of actions, but only d unknown parameters

Optimism for linear bandits

- Same idea
- Estimate θ
- Build confidence intervals
- Play the action that maximizes an upper confidence bound

Least squares estimation

- $A_1, \ldots, A_t \in \mathbb{R}^d$
- $X_1, \ldots, X_t \in \mathbb{R}$
- (regularized) Least squares estimator

$$\hat{\theta}_t = \mathrm{argmin}_{\hat{\theta}} \sum_{t=1}^{n} \left(X_t - \langle A_t, \hat{\theta} \rangle \right)^2 + \lambda \|\hat{\theta}\|_2^2$$

- **Exercise** Show that $\hat{\theta}_t = G_t^{-1} S_t$

$$G_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top \qquad S_t = \sum_{s=1}^{t} A_s X_s$$

Least squares estimation

- $A_1, \ldots, A_t \in \mathbb{R}^d$ and $X_1, \ldots, X_t \in \mathbb{R}$ and $\hat{\theta}_t = G_t^{-1} S_t$

$$G_t = \lambda I + \sum_{s=1}^t A_s A_s^\top \qquad S_t = \sum_{s=1}^t A_s X_s$$

- When $\lambda = 0$
- **Unbiased** $\mathbb{E}[\hat{\theta}_t] = \theta$
- **Variance** $\mathbb{E}[\langle x, \hat{\theta}_t - \theta \rangle^2] = \|x\|_{G_t^{-1}}^2 = x^\top G_t^{-1} x$

Least squares estimation

- **Subtle issue** Fixed design or sequential design
- When $A_1, \ldots, A_t$ are chosen in advance,

$$\mathbb{P}\left(\langle x, \hat{\theta} - \theta \rangle \geq \sqrt{2 \|x\|^2_{G_t^{-1}} \log(1/\delta)}\right) \leq \delta$$

- Easy proof (exercise!)
- Result is **not true** when $A_1, \ldots, A_t$ are chosen sequentially

$$\mathbb{P}\left(\langle x, \hat{\theta} - \theta \rangle \geq \sqrt{2d \|x\|^2_{G_t^{-1}} \log(1/\delta)}\right) \lesssim \delta$$

- More difficult proof

UCB for contextual linear bandits

- Observe $\mathcal{A}_t$
- Choose $A_t = \operatorname{argmax}_{a \in \mathcal{A}_t} \langle \hat{\theta}_t, a \rangle + \beta_t \|a\|_{G_{t-1}^{-1}}$

$$\beta_t \approx \sqrt{d \log(t)}$$

- Observe X_t and update least squares estimator