Multi-armed Bandits (MAB)

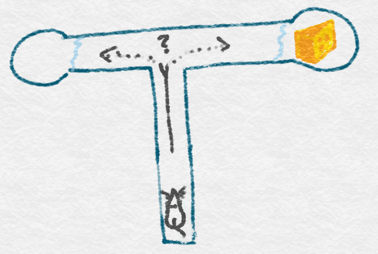
VG441 Cong Shi

Overview

- What are bandits, and why you should care
- Finite-armed stochastic bandits
  - Explore-Then-Commit (ETC) Algorithm
  - Upper Confidence Bound (UCB) Algorithm
  - Lower Bound
- Finite-armed adversarial bandits

What's in a name? A tiny bit of history

First bandit algorithm proposed by Thompson (1933)



Bush and Mosteller (1953) were interested in how mice behaved in a T-maze

Applications

- Clinical trials/dose discovery
- Recommendation systems (movies/news/etc)
- Advertisement placement
- A/B testing
- Dynamic pricing (eg., for Amazon products)
- Ranking (eg., for search)
- Resource allocation

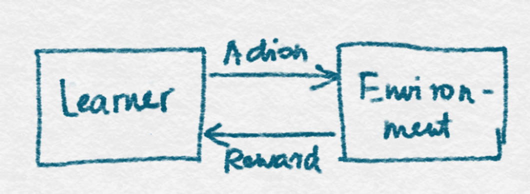
- They isolate an important component of reinforcement learning: exploration-vs-exploitation

Finite-armed bandits

- K actions
- n rounds
- In each round t the **learner** chooses an action

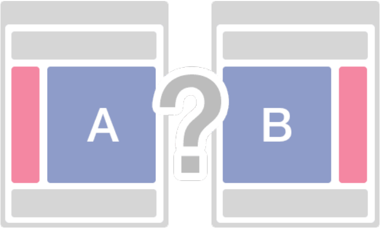
$$A_t \in \{1, 2, \ldots, K\}\,.$$

- Observes **reward** $X_t \sim P_{A_t}$ where $P_1, P_2, \ldots, P_K$ are **unknown** distributions (Gaussian or subgaussian)



Example: A/B testing

- Business wants to optimize their webpage
- Actions correspond to ‘A’ and ‘B’ (two arms)
- Users arrive at webpage sequentially
- Algorithm chooses either ‘A’ or ‘B’ (pulling an arm)
- Receives activity feedback (click as the reward)



Measuring performance – the **regret**

- Let μ_i be the mean reward of distribution P_i
- $\mu^* = \max_i \mu_i$ is the maximum mean
- The (expected) **regret** is

$$R_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$

Measuring performance – the **regret**

- Let μ_i be the mean reward of distribution P_i
- $\mu^* = \max_i \mu_i$ is the maximum mean
- The (expected) **regret** is

$$R_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$

- A reasonable policy for which the regret should be ($R_n = o(n)$)
- Of course we would like to make it as 'small as possible'

Measuring performance – the **regret**

Let $\Delta_i = \mu^* - \mu_i$ be the **suboptimality gap** for the ith arm

Let $T_i(n)$ be the number of times arm i is played over all n rounds

Key decomposition lemma: $R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$

Measuring performance – the **regret**

Let $\Delta_i = \mu^* - \mu_i$ be the **suboptimality gap** for the ith arm

Let $T_i(n)$ be the number of times arm i is played over all n rounds

Key decomposition lemma: $R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$

Proof Let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | A_1, X_1, \ldots, X_{t-1}, A_t]$

$$R_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right] = n\mu^* - \sum_{t=1}^{n} \mathbb{E}[\mathbb{E}_t[X_t]] = n\mu^* - \sum_{t=1}^{n} \mathbb{E}[\mu_{A_t}]$$

$$= \sum_{t=1}^{n} \mathbb{E}[\Delta_{A_t}] = \mathbb{E}\left[\sum_{t=1}^{n} \sum_{i=1}^{K} \mathbb{1}(A_t = i)\Delta_i\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{K} \Delta_i \sum_{t=1}^{n} \mathbb{1}(A_t = i)\right] = \mathbb{E}\left[\sum_{i=1}^{K} \Delta_i T_i(n)\right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$$

A simple policy: Explore-Then-Commit

1 Choose each action m times

2 Find the empirically best action $I \in \{1, 2, \ldots, K\}$ (i.e., the action I gives the largest average reward over m items)

3 Choose $A_t = I$ for all remaining $(n - mK)$ rounds

A simple policy: Explore-Then-Commit

1 Choose each action m times

2 Find the empirically best action $I \in \{1, 2, \ldots, K\}$ (i.e., the action I gives the largest average reward over m items)

3 Choose $A_t = I$ for all remaining $(n - mK)$ rounds

In order to analyse this policy we need to bound the probability of committing to a suboptimal action
Need probability tools: concentration inequalities.

A crash course in concentration

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of independent and identically distributed random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 < \infty$

$$\text{empirical mean} = \hat{\mu}_n = \frac{1}{n} \sum_{t=1}^{n} Z_t$$

How close is $\hat{\mu}_n$ to μ?

A crash course in concentration

$$\text{empirical mean} = \hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} Z_t$$

How close is $\hat{\mu}_n$ to μ?

Classical statistics says:

1. (law of large numbers) $\lim_{n\to\infty} \hat{\mu}_n = \mu$ almost surely
2. (central limit theorem) $\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$
3. (Chebyshev's inequality) $\mathbb{P}\left(|\hat{\mu}_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$ $\left(\mathbb{V}(\hat{\mu}_n) = \frac{\sigma^2}{n}\right)$

A crash course in concentration

$$\text{empirical mean} = \hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} Z_t$$

How close is $\hat{\mu}_n$ to μ?

Classical statistics says:

1. (law of large numbers) $\lim_{n\to\infty} \hat{\mu}_n = \mu$ almost surely
2. (central limit theorem) $\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$
3. (Chebyshev's inequality) $\mathbb{P}\left(|\hat{\mu}_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$ $\left(\mathbb{V}(\hat{\mu}_n) = \frac{\sigma^2}{n}\right)$

Basic probability inequality (R.V. X with finite mean and variance):

1. (Markov's inequality) $\mathbb{P}\left(|X| \geq \varepsilon\right) \leq \frac{\mathbb{E}(|X|)}{\varepsilon}$.
2. (Chebyshev's inequality) $\mathbb{P}\left(|X - \mathbb{E}(X)| \geq \varepsilon\right) \leq \frac{\mathbb{V}(X)}{\varepsilon^2}$

A crash course in concentration

$$\text{empirical mean} = \hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} Z_t$$

How close is $\hat{\mu}_n$ to μ?

Classical statistics says:

1. (law of large numbers) $\lim_{n\to\infty} \hat{\mu}_n = \mu$ almost surely
2. (central limit theorem) $\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$
3. (Chebyshev's inequality) $\mathbb{P}\left(|\hat{\mu}_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$ $\left(\mathbb{V}(\hat{\mu}_n) = \frac{\sigma^2}{n}\right)$

Basic probability inequality (R.V. X with finite mean and variance):

1. (Markov's inequality) $\mathbb{P}\left(|X| \geq \varepsilon\right) \leq \frac{\mathbb{E}(|X|)}{\varepsilon}$.
2. (Chebyshev's inequality) $\mathbb{P}\left(|X - \mathbb{E}(X)| \geq \varepsilon\right) \leq \frac{\mathbb{V}(X)}{\varepsilon^2}$

We need something nonasymptotic and stronger than Chebyshev's (Not possible without assumptions)

A crash course in concentration

Random variable Z is σ-subgaussian if for all $\lambda \in \mathbb{R}$,

$$M_Z(\lambda) \doteq \mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\lambda^2\sigma^2/2\right),$$

where $M_Z(\lambda)$ is known as the moment generating function.

A crash course in concentration

Random variable Z is σ-subgaussian if for all $\lambda \in \mathbb{R}$,

$$M_Z(\lambda) \doteq \mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\lambda^2\sigma^2/2\right),$$

where $M_Z(\lambda)$ is known as the moment generating function.

- Which distributions are σ-subgaussian? Gaussian, Bernoulli, bounded support.
- And not: exponential, power law

A crash course in concentration

Random variable Z is σ-subgaussian if for all $\lambda \in \mathbb{R}$,

$$\boxed{M_Z(\lambda) \doteq \mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\lambda^2\sigma^2/2\right),}$$

where $M_Z(\lambda)$ is known as the moment generating function.

- Which distributions are σ-subgaussian? Gaussian, Bernoulli, bounded support.
- And not: exponential, power law

Lemma If $Z, Z_1, \ldots, Z_n$ are independent and σ-subgaussian, then

- aZ is $|a|\sigma$-subgaussian for any $a \in \mathbb{R}$
- $\sum_{t=1}^n Z_t$ is $\sqrt{n}\sigma$-subgaussian
- $\hat{\mu}_n$ is $n^{-1/2}\sigma$-subgaussian

A crash course in concentration

Lemma(Tail bound of subgaussian random variable)

- If X is a σ-subgaussian, for any $\varepsilon > 0$, $\mathbb{P}(X \geq \varepsilon) \leq \exp(-\frac{\varepsilon^2}{2\sigma^2})$.

A crash course in concentration

Lemma(Tail bound of subgaussian random variable)

- If X is a σ-subgaussian, for any $\varepsilon > 0$, $\mathbb{P}(X \geq \varepsilon) \leq \exp(-\frac{\varepsilon^2}{2\sigma^2})$.

Proof We use $\boxed{\textbf{Chernoff's method}}$. Let $\varepsilon > 0$ and $\lambda = \varepsilon/\sigma^2$.

$$\begin{aligned}
\mathbb{P}(X \geq \varepsilon) &= \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda\varepsilon)) \\
&\leq \frac{\mathbb{E}[\exp(\lambda X)]}{\exp(\lambda\varepsilon)} && \text{(Markov's)} \\
&\leq \exp\left(\sigma^2\lambda^2/2 - \lambda\varepsilon\right) && \text{(X is subgaussian)} \\
&= \exp\left(-\varepsilon^2/(2\sigma^2)\right)
\end{aligned}$$

A crash course in concentration

Theorem If $Z_1, \ldots, Z_n$ are independent and σ-subgaussian, then

$$\mathbb{P}\left(\hat{\mu}_n - \mu \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\right) \leq \delta$$

Proof $\hat{\mu}_n - \mu$ is a $\sigma/\sqrt{n}$-subgaussian random variable and thus

$$\mathbb{P}\left(\hat{\mu}_n - \mu \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

Setting $\exp(-\frac{n\varepsilon^2}{2\sigma^2}) = \delta$ and solving for ε.

A crash course in concentration

Theorem If $Z_1, \ldots, Z_n$ are independent and σ-subgaussian, then

$$\mathbb{P}\left(\hat{\mu}_n - \mu \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\right) \leq \delta$$

Proof $\hat{\mu}_n - \mu$ is a $\sigma/\sqrt{n}$-subgaussian random variable and thus

$$\mathbb{P}\left(\hat{\mu}_n - \mu \geq \varepsilon\right) \leq \exp(-\frac{n\varepsilon^2}{2\sigma^2})$$

Setting $\boxed{\exp(-\frac{n\varepsilon^2}{2\sigma^2}) = \delta}$ and solving for ε.

Corollary If $Z_1, \ldots, Z_n$ are independent and σ-subgaussian, then

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

A crash course in concentration

- Comparing Chebyshev's w. subgaussian bound:

$$\textbf{Chebyshev's:} \quad \mathbb{P}\left(\hat{\mu}_n - \mu \geq \sqrt{\frac{\sigma^2}{n\delta}}\right) \leq \delta$$

$$\textbf{Subgaussian:} \quad \mathbb{P}\left(\hat{\mu}_n - \mu \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\right) \leq \delta$$

A crash course in concentration

- Comparing Chebyshev's w. subgaussian bound:

$$\textbf{Chebyshev's:} \quad \mathbb{P}\left(\hat{\mu}_n - \mu \geq \sqrt{\frac{\sigma^2}{n\delta}}\right) \leq \delta$$

$$\textbf{Subgaussian:} \quad \mathbb{P}\left(\hat{\mu}_n - \mu \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\right) \leq \delta$$

- Typically $\delta \ll 1/n$ in our use-cases. Then Chebyshev's inequality is too loose since $\sqrt{\frac{\sigma^2}{n\delta}}$ is too large.

From now on, we will assume that reward distribution associated with each arm is 1-subgaussian (but with different means)

Analysing Explore-Then-Commit

- **Exploration phase:** Chooses each arm m times
- **Exploitation phase:** Then commits to the arm with the largest empirical reward

Analysing Explore-Then-Commit

- **Exploration phase:** Chooses each arm m times
- **Exploitation phase:** Then commits to the arm with the largest empirical reward

- **Standard convention:** Assume $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$
- Means that first arm is optimal
- Algorithms are symmetric and do not know this fact

Analysing Explore-Then-Commit

- **Exploration phase:** Chooses each arm m times
- **Exploitation phase:** Then commits to the arm with the largest empirical reward

- **Standard convention:** Assume $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$
- Means that first arm is optimal
- Algorithms are symmetric and do not know this fact

- We consider only $K = 2$

Analysing Explore-Then-Commit

Step 1 Let $\hat{\mu}_i$ be the average reward of i-th arm (for $i \in \{1, 2\}$) after the exploration phase

The algorithm commits to the wrong arm if

$$\hat{\mu}_2 \geq \hat{\mu}_1 \Leftrightarrow \hat{\mu}_2 - \mu_2 + \mu_1 - \hat{\mu}_1 \geq \Delta = \mu_1 - \mu_2$$

Observation $\underbrace{\hat{\mu}_2 - \mu_2}_{\sqrt{1/m}-\text{subgaussian}} + \underbrace{\mu_1 - \hat{\mu}_1}_{\sqrt{1/m}-\text{subgaussian}}$ is $\sqrt{2/m}$-subgaussian with zero-mean

Analysing Explore-Then-Commit

Step 1 The algorithm commits to the wrong arm if

$$\hat{\mu}_2 \geq \hat{\mu}_1 \Leftrightarrow \hat{\mu}_2 - \mu_2 + \mu_1 - \hat{\mu}_1 \geq \Delta = \mu_1 - \mu_2$$

Observation $\hat{\mu}_2 - \mu_2 + \mu_1 - \hat{\mu}_1$ is $\sqrt{2/m}$-subgaussian with zero-mean

Analysing Explore-Then-Commit

Step 1 The algorithm commits to the wrong arm if

$$\hat{\mu}_2 \geq \hat{\mu}_1 \Leftrightarrow \hat{\mu}_2 - \mu_2 + \mu_1 - \hat{\mu}_1 \geq \Delta = \mu_1 - \mu_2$$

Observation $\hat{\mu}_2 - \mu_2 + \mu_1 - \hat{\mu}_1$ is $\sqrt{2/m}$-subgaussian with zero-mean

Step 2 The regret is

$$\begin{aligned} R_n = \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}\right] &= \mathbb{E}\left[\sum_{t=1}^{2m} \Delta_{A_t}\right] + \mathbb{E}\left[\sum_{t=2m+1}^{n} \Delta_{A_t}\right] \\ &= m\Delta + (n-2m)\Delta\mathbb{P}\left(\text{commit to the wrong arm}\right) \\ &= m\Delta + (n-2m)\Delta\mathbb{P}\left(\hat{\mu}_2 - \mu_2 + \mu_1 - \hat{\mu}_1 \geq \Delta\right) \\ &\leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \end{aligned}$$

The last inequality is because *if X is a σ-subgaussian, for any $\varepsilon > 0$,* $\mathbb{P}\left(X \geq \varepsilon\right) \leq \exp(-\frac{\varepsilon^2}{2\sigma^2})$ ($\varepsilon = \Delta$ and $\sigma = \sqrt{2/m}$).

Analysing Explore-Then-Commit

$$R_n \le \underbrace{m\Delta}_{\text{(A)}} + \underbrace{n\Delta \exp(-m\Delta^2/4)}_{\text{(B)}}$$

(A) is monotone increasing in m while (B) is monotone decreasing in m

Exploration/Exploitation Trade-off Exploring too much (m large) then (A) is big, while exploring too little makes (B) large

Bound minimised by $m = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil$ leading to

$$R_n \le \Delta + \frac{4}{\Delta} \log\left(\frac{n\Delta^2}{4}\right) + \frac{4}{\Delta},$$

noting that due to ceiling function in m: $(A) \le (1 + \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right))\Delta$.

Analysing Explore-Then-Commit

Last slide: $R_n \leq \Delta + \frac{4}{\Delta} \log\left(\frac{n\Delta^2}{4}\right) + \frac{4}{\Delta}$

What happens when Δ is very small? (R_n can be unbounded)

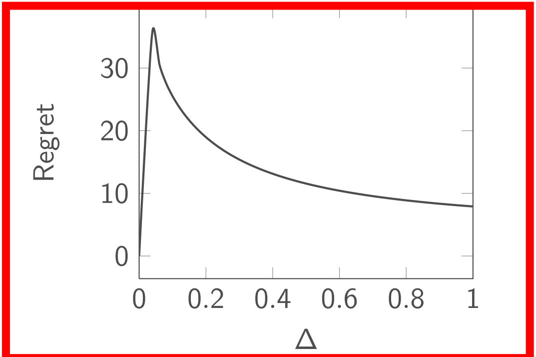
Analysing Explore-Then-Commit

Last slide: $R_n \leq \Delta + \frac{4}{\Delta}\log\left(\frac{n\Delta^2}{4}\right) + \frac{4}{\Delta}$

What happens when Δ is very small? (R_n can be unbounded)
A natural correction:

$$R_n \leq \min\left\{n\Delta,\ \Delta + \frac{4}{\Delta}\log\left(\frac{n\Delta^2}{4}\right) + \frac{4}{\Delta}\right\}$$

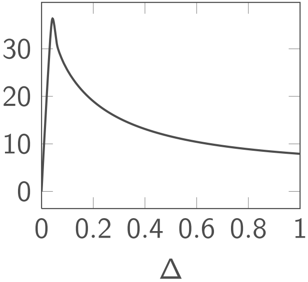
Illustration of R_n with $n = 1000$.



Analysing Explore-Then-Commit

Does this figure make sense? Why is the regret largest when Δ is small, but not too small?

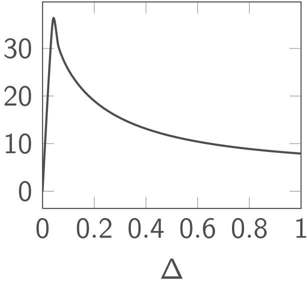
$$R_n \leq \min\left\{n\Delta,\, \Delta + \frac{4}{\Delta}\log\left(\frac{n\Delta^2}{4}\right) + \frac{4}{\Delta}\right\}$$



Analysing Explore-Then-Commit

Does this figure make sense? Why is the regret largest when Δ is small, but not too small?

$$R_n \leq \min\left\{n\Delta,\ \Delta + \frac{4}{\Delta}\log\left(\frac{n\Delta^2}{4}\right) + \frac{4}{\Delta}\right\}$$



Small Δ makes identification of the best arm hard, but cost of failure (of identification) is low

Large Δ makes the cost of failure high, but identification becomes easy

Worst case is when $\Delta \approx \sqrt{1/n}$ with $R_n \approx \sqrt{n}$

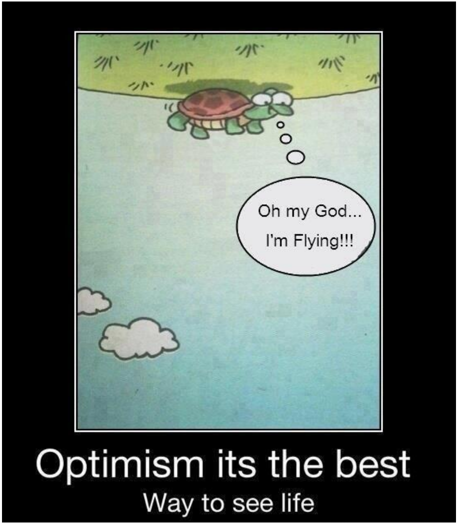
Limitations of Explore-Then-Commit

- Recall that $m = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil$
- Need advance knowledge of the unknown horizon length n
- Optimal tuning depends on unknown $\Delta = \mu_1 - \mu_2$

Limitations of Explore-Then-Commit

- Recall that $m = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil$
- Need advance knowledge of the unknown horizon length n
- Optimal tuning depends on unknown $\Delta = \mu_1 - \mu_2$
- Better approaches now exist, but Explore-Then-Commit is often a good place to start when analyzing a bandit problem since it captures *exploration-exploitation trade-off*

Optimism principle



Informal illustration

Visiting a new region

Shall I try local cuisine?

Optimist: Yes!

Pessimist: No!

Optimism leads to exploration, pessimism prevents it

Exploration is necessary, but how much?

Optimism principle

- Let $\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{1}(A_s = i) X_s$ be the empirical mean reward of i-th arm at time t

Optimism principle

- Let $\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{1}(A_s = i) X_s$ be the empirical mean reward of i-th arm at time t
- Optimistic estimate of the mean of arm = 'largest value it could plausibly be'

Optimism principle

- Let $\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{1}(A_s = i) X_s$ be the empirical mean reward of i-th arm at time t
- Optimistic estimate of the mean of arm = 'largest value it could plausibly be'
- Formalise the intuition using confidence intervals ($\sigma = 1$)

$$\mathbb{P}\left(\hat{\mu}_n - \mu \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\right) \leq \delta$$

- Suggests

$$\text{optimistic estimate} = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$$

- $\delta \in (0, 1)$ determines the level of optimism

Upper confidence bound algorithm

1. Choose each action once
2. Choose the action maximising

$$A_t = \text{argmax}_i \, \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t^3)}{T_i(t-1)}}$$

3. **Goto** 2

Upper confidence bound algorithm

1 Choose each action once

2 Choose the action maximising

$$A_t = \text{argmax}_i \, \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t^3)}{T_i(t-1)}}$$

3 **Goto** 2

Corresponds to $\delta = 1/t^3$. This is quite a conservative choice (more on this later)

Upper confidence bound algorithm

1 Choose each action once

2 Choose the action maximising

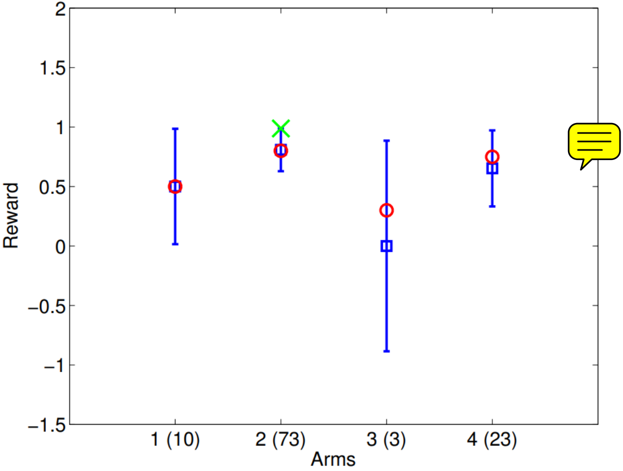
$$A_t = \text{argmax}_i \, \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t^3)}{T_i(t-1)}}$$

3 **Goto** 2

Corresponds to $\delta = 1/t^3$. This is quite a conservative choice (more on this later)

Algorithm does not depend on horizon n (it is **anytime**)

Upper confidence bound algorithm



- Red circle: true mean, Blue rectangle: empirical mean reward.
- (10), (73), (3), (23): number of pulls (a larger number of pulls makes the true and empirical mean closer).

Why UCB?

- A suboptimal arm can only be played if its upper confidence bound is larger than the upper confidence bound of the optimal arm, which in turn is larger than the mean of the optimal arm.

Why UCB?

- A suboptimal arm can only be played if its upper confidence bound is larger than the upper confidence bound of the optimal arm, which in turn is larger than the mean of the optimal arm.

- However,this cannot happen too often because by playing a few more times of a suboptimal arm, its upper confidence bound will be close to its true mean. Thus, it will eventually fall below the upper confidence bound of the optimal arm.

Why UCB?

- A suboptimal arm can only be played if its upper confidence bound is larger than the upper confidence bound of the optimal arm, which in turn is larger than the mean of the optimal arm.

- However,this cannot happen too often because by playing a few more times of a suboptimal arm, its upper confidence bound will be close to its true mean. Thus, it will eventually fall below the upper confidence bound of the optimal arm.

- UCB:

$$A_t = \text{argmax}_i \, \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t^3)}{T_i(t-1)}}$$

- An algorithm should explore arms more often if they are
  1. either promising because $\hat{\mu}_i(t-1)$ is large
  2. or not well explored because $T_i(t-1)$ is small

Regret of UCB

Theorem The regret of UCB is at most

$$R_n = O\left(\sum_{i:\Delta_i>0}\left(\Delta_i + \frac{\log(n)}{\Delta_i}\right)\right)$$

Furthermore,

$$R_n = O\left(\sqrt{Kn\log(n)}\right),$$

where K is the number of arms and n is the time horizon length.

Bounds of the first kind are called **problem dependent** or **instance dependent**, which depends on $\Delta_i = \mu_1 - \mu_i$

Bounds like the second are called **distribution free** or **worst case**

Regret analysis

Rewrite the regret $R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$

Only need to show that $\mathbb{E}[T_i(n)]$ is not too large for suboptimal arms

Regret analysis

Key insight Arm i is only played if its **index** is larger than the index of the optimal arm

$$\gamma_i(t-1) = \underbrace{\hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t^3)}{T_i(t-1)}}}_{\text{index of arm } i \text{ in round } t}$$

Regret analysis

Key insight Arm i is only played if its **index** is larger than the index of the optimal arm

$$\gamma_i(t-1) = \underbrace{\hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t^3)}{T_i(t-1)}}}_{\text{index of arm } i \text{ in round } t}$$

A suboptimal arm $i \neq 1$ is played implies that

1. either $\gamma_i(t-1) \geq \mu_1$ (index of arm i is larger than the mean of optimal arm)
2. or $\gamma_1(t-1) \leq \mu_1$ (index of arm 1 is smaller than its true mean)

Otherwise, we have $\gamma_i(t-1) \leq \mu_1 \leq \gamma_1(t-1)$: arm 1 should be played since

Both events are unlikely after a sufficiently number of plays.

Regret analysis

To make this intuition a reality we decompose the "pull-count" for the i-th arm ($i \neq 1$)

$$\begin{aligned}
\mathbb{E}[T_i(n)] &= \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}(A_t = i)\right] = \sum_{t=1}^{n} \mathbb{P}\left(A_t = i\right) \\
&= \sum_{t=1}^{n} \mathbb{P}\left(A_t = i \text{ and } \left(\gamma_1(t-1) \leq \mu_1 \text{ or } \gamma_i(t-1) \geq \mu_1\right)\right) \\
&\leq \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\gamma_1(t-1) \leq \mu_1\right)}_{\text{index of opt. arm too small?}} + \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(A_t = i \text{ and } \gamma_i(t-1) \geq \mu_1\right)}_{\text{index of subopt. arm large?}}
\end{aligned}$$

Regret analysis

We want to show that $\mathbb{P}(\gamma_1(t-1) \leq \mu_1)$ is small

Tempting to use the concentration theorem...

$$\mathbb{P}(\gamma_1(t-1) \leq \mu_1) = \mathbb{P}\left(\hat{\mu}_1(t-1) + \sqrt{\frac{2\log(t^3)}{T_i(t-1)}} \leq \mu_1\right) \overset{?}{\leq} \frac{1}{t^3}$$

What's wrong with this?

Regret analysis

We want to show that $\mathbb{P}(\gamma_1(t-1) \le \mu_1)$ is small

Tempting to use the concentration theorem...

$$\mathbb{P}(\gamma_1(t-1) \le \mu_1) = \mathbb{P}\left(\hat{\mu}_1(t-1) + \sqrt{\frac{2\log(t^3)}{T_i(t-1)}} \le \mu_1\right) \overset{?}{\le} \frac{1}{t^3}$$

What's wrong with this? $T_i(t-1)$ is a random variable but not a number! Use union bound $\Pr(\cup_{s=1}^{t-1} A_s) \le \sum_{s=1}^{t-1} \Pr(A_s)$

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}_1(t-1) + \sqrt{\frac{2\log(t^3)}{T_i(t-1)}} \le \mu_1\right) &\le \mathbb{P}\left(\exists s \le t-1 : \hat{\mu}_{1,s} + \sqrt{\frac{2\log(t^3)}{s}} \le \mu_1\right) \\ &\le \sum_{s=1}^{t-1} \mathbb{P}\left(\hat{\mu}_{1,s} + \sqrt{\frac{2\log(t^3)}{s}} \le \mu_1\right) \\ &\le \sum_{s=1}^{t-1} \frac{1}{t^3} \le \frac{1}{t^2}\,. \qquad (\delta = 1/t^3)\end{aligned}$$

Regret analysis

$$\begin{aligned}
\sum_{t=1}^{n} \mathbb{P}\left(A_t = i \text{ and } \gamma_i(t-1) \geq \mu_1\right) &= \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}(A_t = i \text{ and } \gamma_i(t-1) \geq \mu_1)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\left(A_t = i \text{ and } \hat{\mu}_i(t-1) + \sqrt{\frac{6\log(t)}{T_i(t-1)}} \geq \mu_1\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\left(A_t = i \text{ and } \hat{\mu}_i(t-1) + \sqrt{\frac{6\log(n)}{T_i(t-1)}} \geq \mu_1\right)\right] \qquad (t \leq n)
\end{aligned}$$

Regret analysis

$$
\begin{aligned}
&\sum_{t=1}^{n} \mathbb{P}\left(A_t = i \text{ and } \gamma_i(t-1) \geq \mu_1\right) \\
\leq &\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}(A_t = i \text{ and } \hat{\mu}_i(t-1) + \sqrt{\frac{6\log(n)}{T_i(t-1)}} \geq \mu_1)\right] \\
\leq &\mathbb{E}\left[\sum_{s=1}^{n} \mathbb{1}(\hat{\mu}_{i,s} + \sqrt{\frac{6\log(n)}{s}} \geq \mu_1)\right] \\
&\qquad\qquad \text{(For each possible } T_i(t-1) = s \text{ and } s = 1, \ldots, n\text{)} \\
= &\sum_{s=1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} + \sqrt{\frac{6\log(n)}{s}} \geq \mu_1\right)
\end{aligned}
$$

Regret analysis

Let $u = \dfrac{24\log(n)}{\Delta_i^2}$. Then we decompose time periods into $[1, u]$ and $[u+1, n]$:

$$\sum_{s=1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} + \sqrt{\frac{6\log(n)}{s}} \geq \mu_1\right) \leq u + \sum_{s=u+1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} + \sqrt{\frac{6\log(n)}{s}} \geq \mu_1\right)$$

Regret analysis

Let $u = \frac{24\log(n)}{\Delta_i^2}$. Then we decompose time periods into $[1, u]$ and $[u+1, n]$:

$$\sum_{s=1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} + \sqrt{\frac{6\log(n)}{s}} \geq \mu_1\right) \leq u + \sum_{s=u+1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} + \sqrt{\frac{6\log(n)}{s}} \geq \mu_1\right)$$

Choose u large enough so that for any $s > u$, $\sqrt{\frac{6\log(n)}{s}} \leq \frac{\Delta_i}{2}$ ($u = \frac{24\log(n)}{\Delta_i^2}$). Then we have

$$\begin{aligned}
\hat{\mu}_{i,s} \geq \mu_1 - \sqrt{\frac{6\log(n)}{s}} \quad &\Rightarrow \quad \hat{\mu}_{i,s} - \mu_i \geq \mu_1 - \mu_i - \sqrt{\frac{6\log(n)}{s}} \\
&\Rightarrow \quad \hat{\mu}_{i,s} - \mu_i \geq \Delta_i - \frac{\Delta_i}{2} \\
&\Rightarrow \quad \hat{\mu}_{i,s} - \mu_i \geq \frac{\Delta_i}{2}
\end{aligned}$$

Regret analysis

Let $u = \dfrac{24\log(n)}{\Delta_i^2}$. Then

$$\sum_{s=1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} + \sqrt{\frac{6\log(n)}{s}} \geq \mu_1\right) \leq u + \sum_{s=u+1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} + \sqrt{\frac{6\log(n)}{s}} \geq \mu_1\right)$$

$$\leq u + \sum_{s=u+1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} \geq \mu_i + \frac{\Delta_i}{2}\right)$$

$$\leq u + \sum_{s=u+1}^{\infty} \exp\left(-\frac{s\Delta_i^2}{8}\right)$$

$$\left(\textstyle\sum_{s=u+1}^{\infty} \exp\left(-\frac{s\Delta_i^2}{8}\right) \leq 1 + \int_{s=u}^{\infty} \exp\left(-\frac{s\Delta_i^2}{8}\right) \mathrm{d}s\right)$$

$$\leq u + 1 + \frac{8}{\Delta_i^2}\,.$$

Regret analysis

Combining the two parts we have

$$\mathbb{E}[T_i(n)] \le \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\gamma_1(t-1) \le \mu_1\right)}_{\text{index of opt. arm too small?}} + \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(A_t = i \text{ and } \gamma_i(t-1) \ge \mu_1\right)}_{\text{index of subopt. arm large?}}$$

$$\le \sum_{t=1}^{n} \frac{1}{t^2} + 1 + u + \frac{8}{\Delta_i^2}$$

$$\left(u = \frac{24\log(n)}{\Delta_i^2},\ \sum_{t=1}^{n} \frac{1}{t^2} \le 1 + \int_{t=1}^{\infty} \frac{1}{t^2}\mathrm{d}t\right)$$

$$\le 3 + \frac{8}{\Delta_i^2} + \frac{24\log(n)}{\Delta_i^2}$$

Regret analysis

Combining the two parts we have

$$\mathbb{E}[T_i(n)] \leq \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\gamma_1(t-1) \leq \mu_1\right)}_{\text{index of opt. arm too small?}} + \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(A_t = i \text{ and } \gamma_i(t-1) \geq \mu_1\right)}_{\text{index of subopt. arm large?}}$$

$$\leq \sum_{t=1}^{n} \frac{1}{t^2} + 1 + u + \frac{8}{\Delta_i^2}$$

$$\left(u = \frac{24\log(n)}{\Delta_i^2},\ \sum_{t=1}^{n} \frac{1}{t^2} \leq 1 + \int_{t=1}^{\infty} \frac{1}{t^2}\mathrm{d}t\right)$$

$$\leq 3 + \frac{8}{\Delta_i^2} + \frac{24\log(n)}{\Delta_i^2}$$

So the regret is bounded by (instance dependent bound)

$$R_n = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)] \leq \sum_{i:\Delta_i>0} \left(3\Delta_i + \frac{8}{\Delta_i} + \frac{24\log(n)}{\Delta_i}\right)$$

$$= O\left(\sum_{i:\Delta_i>0} \left(\Delta_i + \frac{\log(n)}{\Delta_i}\right)\right)$$

Distribution free bounds

Let $\Delta > 0$ be some constant to be chosen later

$$
\begin{aligned}
R_n &= \sum_{i:\Delta_i \leq \Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i > \Delta} \Delta_i \mathbb{E}[T_i(n)] \\
&\leq n\Delta + \sum_{i:\Delta_i > \Delta} \Delta_i \mathbb{E}[T_i(n)] \qquad \left(\textstyle\sum_{i:\Delta_i \leq \Delta} T_i(n) \leq \sum_i T_i(n) = n\right) \\
&\lesssim n\Delta + \sum_{i:\Delta_i > \Delta} \left(\Delta_i + \frac{\log(n)}{\Delta_i}\right) \\
&\lesssim \underbrace{n\Delta + \frac{K\log(n)}{\Delta}}_{\Delta = \sqrt{K\log(n)/n}} + \sum_{i=1}^{K} \Delta_i \\
&\lesssim \sqrt{nK\log(n)} + \sum_{i=1}^{K} \Delta_i
\end{aligned}
$$

Distribution free bounds

Let $\Delta > 0$ be some constant to be chosen later

$$\begin{aligned}
R_n &= \sum_{i:\Delta_i \le \Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i > \Delta} \Delta_i \mathbb{E}[T_i(n)] \\
&\le n\Delta + \sum_{i:\Delta_i > \Delta} \Delta_i \mathbb{E}[T_i(n)] \qquad (\textstyle\sum_{i:\Delta_i \le \Delta} T_i(n) \le \sum_i T_i(n) = n) \\
&\lesssim n\Delta + \sum_{i:\Delta_i > \Delta} (\Delta_i + \frac{\log(n)}{\Delta_i}) \\
&\lesssim \underbrace{n\Delta + \frac{K\log(n)}{\Delta}}_{\Delta = \sqrt{K\log(n)/n}} + \sum_{i=1}^{K} \Delta_i \\
&\lesssim \sqrt{nK\log(n)} + \sum_{i=1}^{K} \Delta_i
\end{aligned}$$

Note that $\sum_{i=1}^{K} \Delta_i$ is unavoidable since each arm needs to be played at least once and this term is negligible when n is large.

Improvements

- The constants in the algorithm/analysis can be improved quite significantly.

$$A_t = \text{argmax}_i \, \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t)}{T_i(t-1)}}$$

- With this choice:

$$\lim_{n\to\infty} \frac{R_n}{\log(n)} = \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}$$

Improvements

- The constants in the algorithm/analysis can be improved quite significantly.

$$A_t = \operatorname{argmax}_i \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t)}{T_i(t-1)}}$$

- With this choice:

$$\lim_{n\to\infty} \frac{R_n}{\log(n)} = \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}$$

- The distribution-free regret is also improvable

$$A_t = \operatorname{argmax}_i \hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log\left(1 + \frac{t}{KT_i(t-1)}\right)}$$

- With this index we save a log factor in the distribution free bound

$$R_n = O(\sqrt{nK})$$

Exercise

- Consider different settings of arms (number of arms K, mean gap Δ_i) and different distributions: uniform, Bernoulli, normal
- Compare Explore-Then-Commit with UCB Algorithm in
    1. Regret as a function of horizon n
    2. Frequency of pulling each arm
    3. Tuning the constant ρ:

$$A_t = \text{argmax}_i \, \hat{\mu}_i(t-1) + \sqrt{\frac{\rho \log(t)}{T_i(t-1)}}$$

Lower bounds

Is the bound $R_n = O(\sqrt{nK})$ optimal in n and K?

1. For worst-case regret for a given policy π: $R_n(\pi) = \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$, where $\mathcal{E}$ denotes the set of K-armed Gaussian bandits with unit variance and means $\mu \in [0, 1]^K$.

Lower bounds

Is the bound $R_n = O(\sqrt{nK})$ optimal in n and K?

1. For worst-case regret for a given policy π: $R_n(\pi) = \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$, where $\mathcal{E}$ denotes the set of K-armed Gaussian bandits with unit variance and means $\mu \in [0, 1]^K$.
2. The minimax regret $R_n^*(\mathcal{E}) = \inf_\pi \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$

Theorem $R_n^*(\mathcal{E}) \geq \sqrt{(K-1)n}/27$: for every policy π and n and $K \leq n + 1$, there exists a K-armed Gaussian bandit ν such that

$$R_n(\pi, \nu) \geq \sqrt{(K-1)n}/27$$

Lower bounds

Is the bound $R_n = O(\sqrt{nK})$ optimal in n and K?

1. For worst-case regret for a given policy π: $R_n(\pi) = \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$, where $\mathcal{E}$ denotes the set of K-armed Gaussian bandits with unit variance and means $\mu \in [0, 1]^K$.
2. The minimax regret $R_n^*(\mathcal{E}) = \inf_\pi \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$

Theorem $R_n^*(\mathcal{E}) \geq \sqrt{(K-1)n}/27$: for every policy π and n and $K \leq n + 1$, there exists a K-armed Gaussian bandit ν such that

$$R_n(\pi, \nu) \geq \sqrt{(K-1)n}/27$$

UCB with $R_n = O(\sqrt{nK})$ is a rate-optimal policy

How to prove a minimax lower bound?

Key idea: reduce the bandit problem into a statistical hypothesis testing problem.

How to prove a minimax lower bound?

Key idea: reduce the bandit problem into a statistical hypothesis testing problem.
Select two bandit problem instances (two sets of K distributions) in such a way that the following two conditions hold simultaneously:

- Competition: A sequence of actions that is good for one bandit is not good for the other (choose two instances far away from each other).
- Similarity: The instances are 'close' enough that a policy interacting with either of the two instances cannot statistically identify the true bandit.

How to prove a minimax lower bound?

Key idea: reduce the bandit problem into a statistical hypothesis testing problem.
Select two bandit problem instances (two sets of K distributions) in such a way that the following two conditions hold simultaneously:

- Competition: A sequence of actions that is good for one bandit is not good for the other (choose two instances far away from each other).
- Similarity: The instances are 'close' enough that a policy interacting with either of the two instances cannot statistically identify the true bandit.

Lower bound: optimize the trade-off between these two opposite goals.

Minimax lower bound

Theorem For every policy π and n and $K \leq n$, there exists a K-armed Gaussian bandit ν such that

$$R_n(\pi, \nu) \geq \sqrt{(K-1)n}/27$$

Proof sketch

- Two bandits: $\nu = (P_i)_{i=1}^K$ and $\nu' = (P'_i)_{i=1}^K$, where $P_i = N(\mu_i, 1)$ and $P'_i = N(\mu'_i, 1)$.

Minimax lower bound

Theorem For every policy π and n and $K \leq n$, there exists a K-armed Gaussian bandit ν such that

$$R_n(\pi, \nu) \geq \sqrt{(K-1)n}/27$$

Proof sketch

- Two bandits: $\nu = (P_i)_{i=1}^K$ and $\nu' = (P'_i)_{i=1}^K$, where $P_i = N(\mu_i, 1)$ and $P'_i = N(\mu'_i, 1)$.
- It suffices to show that for *any policy* π, there exists μ and μ' such that the π incurs regret larger than $\sqrt{Kn}$ on at least one instance:

$$\max(R_n(\pi, \nu), R_n(\pi, \nu')) \geq c\sqrt{Kn},$$

or (since $\max(a, b) \geq (a+b)/2$),

$$R_n(\pi, \nu) + R_n(\pi, \nu') \geq c\sqrt{Kn}.$$

Minimax lower bound

Theorem For every policy π and n and $K \leq n$, there exists a K-armed Gaussian bandit ν such that

$$R_n(\pi, \nu) \geq \sqrt{(K-1)n}/27$$

Proof sketch

- Two bandits: $\nu = (P_i)_{i=1}^K$ and $\nu' = (P'_i)_{i=1}^K$, where $P_i = N(\mu_i, 1)$ and $P'_i = N(\mu'_i, 1)$.
- It suffices to show that for *any policy* π, there exists μ and μ' such that the π incurs regret larger than $\sqrt{Kn}$ on at least one instance:

$$\max(R_n(\pi, \nu), R_n(\pi, \nu')) \geq c\sqrt{Kn},$$

or (since $\max(a, b) \geq (a+b)/2$),

$$R_n(\pi, \nu) + R_n(\pi, \nu') \geq c\sqrt{Kn}.$$

- Choose $\mu = (\Delta, 0, \ldots, 0)$ and

$$R_n(\pi, \nu) = (n - \mathbb{E}_\nu(T_1(n)))\Delta$$

(Δ optimized later)

Minimax lower bound

Theorem For every policy π and n and $K \leq n$, there exists a K-armed Gaussian bandit ν such that

$$R_n(\pi, \nu) \geq \sqrt{(K-1)n}/27$$

Proof sketch

- Choose $\mu = (\Delta, 0, \ldots, 0)$ and $R_n(\pi, \nu) = (n - \mathbb{E}_\nu(T_1(n)))\Delta$
- Let $i = \operatorname{argmin}_{j>1} \mathbb{E}_\nu E(T_j(n))$ (arm explored the least) and $\mathbb{E}_\nu[T_i(n)] \leq n/(K-1)$

Minimax lower bound

Theorem For every policy π and n and $K \leq n$, there exists a K-armed Gaussian bandit ν such that

$$R_n(\pi, \nu) \geq \sqrt{(K-1)n}/27$$

Proof sketch

- Choose $\mu = (\Delta, 0, \ldots, 0)$ and $R_n(\pi, \nu) = (n - \mathbb{E}_\nu(T_1(n)))\Delta$
- Let $i = \text{argmin}_{j>1} \mathbb{E}_\nu E(T_j(n))$ (arm explored the least) and $\mathbb{E}_\nu[T_i(n)] \leq n/(K-1)$
- $\mu' = (\Delta, 0, \ldots, 2\Delta, 0, \ldots, 0)$ (2Δ at the i-th arm, optimal arm):

$$R_n(\pi, \nu') = \Delta \mathbb{E}_{\nu'}(T_1(n)) + \sum_{j \neq 1, i} 2\Delta \mathbb{E}_{\nu'}(T_j(n)) \geq \Delta \mathbb{E}_{\nu'}(T_1(n)).$$

Minimax lower bound

Theorem For every policy π and n and $K \leq n$, there exists a K-armed Gaussian bandit ν such that

$$R_n(\pi, \nu) \geq \sqrt{(K-1)n}/27$$

Proof sketch

- Choose $\mu = (\Delta, 0, \ldots, 0)$ and $R_n(\pi, \nu) = (n - \mathbb{E}_\nu(T_1(n)))\Delta$
- Let $i = \mathrm{argmin}_{j>1} \mathbb{E}_\nu E(T_j(n))$ (arm explored the least) and $\mathbb{E}_\nu[T_i(n)] \leq n/(K-1)$
- $\mu' = (\Delta, 0, \ldots, 2\Delta, 0, \ldots, 0)$ (2Δ at the i-th arm, optimal arm):

$$R_n(\pi, \nu') = \Delta \mathbb{E}_{\nu'}(T_1(n)) + \sum_{j \neq 1, i} 2\Delta \mathbb{E}_{\nu'}(T_j(n)) \geq \Delta \mathbb{E}_{\nu'}(T_1(n)).$$

- Depend on $T_1(n) \gtrless n/2$,
  - If $T_1(n) \leq n/2$, $R_n(\pi, \nu) = (n - \mathbb{E}_\nu(T_1(n)))\Delta \geq \frac{n\Delta}{2}$. Therefore, $R_n(\pi, \nu) \geq \mathbb{P}_\nu(T_1(n) \leq n/2)\frac{n\Delta}{2}$
  - If $T_1(n) \leq n/2$, $R_n(\pi, \nu') \geq \Delta \mathbb{E}_{\nu'}(T_1(n)) \geq \frac{n\Delta}{2}$. Therefore, $R_n(\pi, \nu') \geq \mathbb{P}_{\nu'}(T_1(n) \geq n/2)\frac{n\Delta}{2}$

Minimax lower bound

$$R_n(\pi,\nu) + R_n(\pi,\nu') \geq \frac{n\Delta}{2}(\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2))$$

Need to show $\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2)$ is larger!

Minimax lower bound

$$R_n(\pi,\nu) + R_n(\pi,\nu') \geq \frac{n\Delta}{2}(\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2))$$

Need to show $\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2)$ is larger!

Theorem (Pinsker's inequality) For any two distributions and any event A:

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp(-D(P,Q)),$$

where $D(P,Q) = \int p \log(\frac{p}{q})$ is the Kullback-Leibler (KL) divergence.

Intuition, when P is close to Q, $P(A) + Q(A^c)$ should be large ($P(A) + P(A^c) = 1$)

Minimax lower bound

$$R_n(\pi,\nu) + R_n(\pi,\nu') \geq \frac{n\Delta}{2}(\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2))$$

Need to show $\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2)$ is larger!

Theorem (Pinsker's inequality) For any two distributions and any event A:

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp(-D(P,Q)),$$

where $D(P,Q) = \int p \log(\frac{p}{q})$ is the Kullback-Leibler (KL) divergence.

Intuition, when P is close to Q, $P(A) + Q(A^c)$ should be large $(P(A) + P(A^c) = 1)$

Minimax lower bound

$$R_n(\pi,\nu) + R_n(\pi,\nu') \geq \frac{n\Delta}{2}(\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2))$$

Need to show $\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2)$ is larger!

Theorem (Pinsker's inequality) For any two distributions and any event A:

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp(-D(P,Q)),$$

where $D(P,Q) = \int p \log(\frac{p}{q})$ is the Kullback-Leibler (KL) divergence.

Intuition, when P is close to Q, $P(A) + Q(A^c)$ should be large ($P(A) + P(A^c) = 1$)

Exercise:

$$D(N(\mu_1,\sigma^2), N(\mu_2,\sigma^2)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}$$

Minimax lower bound

$$\begin{aligned} R_n(\pi,\nu) + R_n(\pi,\nu') &\geq \frac{n\Delta}{2}\left(\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2)\right) \\ &\geq \frac{n\Delta}{4}\exp(-D(\mathbb{P}_\nu, \mathbb{P}_{\nu'})) \end{aligned}$$

Minimax lower bound

$$
\begin{aligned}
R_n(\pi,\nu) + R_n(\pi,\nu') &\geq \frac{n\Delta}{2}\left(\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2)\right) \\
&\geq \frac{n\Delta}{4}\exp(-D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}))
\end{aligned}
$$

Theorem (Divergence decomposition)

$$
D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{j=1}^{K} \mathbb{E}_\nu(T_j(n)) D(P_j, P_j')
$$

Minimax lower bound

$$
\begin{aligned}
R_n(\pi,\nu) + R_n(\pi,\nu') &\geq \frac{n\Delta}{2}(\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2)) \\
&\geq \frac{n\Delta}{4}\exp(-D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}))
\end{aligned}
$$

Theorem (Divergence decomposition)

$$
D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{j=1}^{K} \mathbb{E}_\nu(T_j(n)) D(P_j, P_j')
$$

Recall $\mu = (\Delta, 0, \ldots, 0)$ and $\mu' = (\Delta, 0, \ldots, 2\Delta, 0, \ldots, 0)$ (only one entry different):

$$
D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \mathbb{E}_\nu(T_i(n)) D(N(0,1), D(2\Delta, 1)) = \mathbb{E}_\nu(T_i(n)) \frac{(2\Delta)^2}{2} \leq \frac{2n\Delta^2}{K-1}.
$$

$$
R_n(\pi,\nu) + R_n(\pi,\nu') \geq \frac{n\Delta}{4}\exp(-\frac{2n\Delta^2}{K-1})
$$

Minimax lower bound

$$\begin{aligned} R_n(\pi,\nu) + R_n(\pi,\nu') &\geq \frac{n\Delta}{2}(\mathbb{P}_\nu(T_1(n) \leq n/2) + \mathbb{P}_{\nu'}(T_1(n) \geq n/2)) \\ &\geq \frac{n\Delta}{4}\exp(-D(\mathbb{P}_\nu, \mathbb{P}_{\nu'})) \end{aligned}$$

Theorem (Divergence decomposition)

$$D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{j=1}^{K} \mathbb{E}_\nu(T_j(n)) D(P_j, P_j')$$

Recall $\mu = (\Delta, 0, \ldots, 0)$ and $\mu' = (\Delta, 0, \ldots, 2\Delta, 0, \ldots, 0)$ (only one entry different):

$$D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \mathbb{E}_\nu(T_i(n)) D(N(0,1), D(2\Delta, 1)) = \mathbb{E}_\nu(T_i(n)) \frac{(2\Delta)^2}{2} \leq \frac{2n\Delta^2}{K-1}.$$

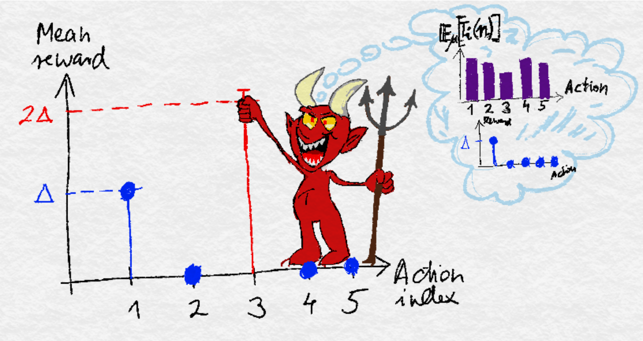
$$R_n(\pi,\nu) + R_n(\pi,\nu') \geq \frac{n\Delta}{4}\exp(-\frac{2n\Delta^2}{K-1})$$

and choose $\Delta = \sqrt{(K-1)/4n}$

Worst case lower bound

Theorem For every policy π and n and $K \leq n + 1$, there exists a K-armed Gaussian bandit ν such that

$$R_n(\pi, \nu) \geq \sqrt{(K-1)n}/27$$



What else is there?

- All kinds of variants of UCB for different noise models: Bernoulli, exponential families, heavy tails, Gaussian with unknown mean and variance,...

What else is there?

- All kinds of variants of UCB for different noise models: Bernoulli, exponential families, heavy tails, Gaussian with unknown mean and variance,...
- Thompson sampling: each round sample mean from posterior for each arm, choose arm with largest

What else is there?

- All kinds of variants of UCB for different noise models: Bernoulli, exponential families, heavy tails, Gaussian with unknown mean and variance,...
- Thompson sampling: each round sample mean from posterior for each arm, choose arm with largest
- All manner of twists on the setup: non-stationarity, delayed rewards, playing multiple arms each round, moving beyond expected regret (high probability bounds)

The adversarial viewpoint

- Replace random rewards with an **adversary**
- At the start of the game the adversary secretly chooses **losses** $\ell_1, \ell_2, \ldots, \ell_n$ where $\ell_t \in [0,1]^K$
- Learner chooses actions A_t:
  - observe and suffers the loss ℓ_{tA_t}
- Regret is

$$R_n = \underbrace{\mathbb{E}\left[\sum_{t=1}^n \ell_{tA_t}\right]}_{\text{learner's loss}} - \underbrace{\min_i \sum_{t=1}^n \ell_{ti}}_{\text{loss of best arm}}$$

- **Mission** Make the regret small, regardless of the adversary

The adversarial viewpoint

- Replace random rewards with an **adversary**
- At the start of the game the adversary secretly chooses **losses** $\ell_1, \ell_2, \ldots, \ell_n$ where $\ell_t \in [0,1]^K$
- Learner chooses actions A_t:
  - observe and suffers the loss ℓ_{tA_t}
- Regret is

$$R_n = \underbrace{\mathbb{E}\left[\sum_{t=1}^n \ell_{tA_t}\right]}_{\text{learner's loss}} - \underbrace{\min_i \sum_{t=1}^n \ell_{ti}}_{\text{loss of best arm}}$$

- **Mission** Make the regret small, regardless of the adversary
- There exists an algorithm such that

$$R_n \leq 2\sqrt{Kn}$$

Why this regret definition?

- The regret is with respect to the loss of the best arm

$$R_n = \underbrace{\mathbb{E}\left[\sum_{t=1}^{n} \ell_{tA_t}\right]}_{\text{learner's loss}} - \underbrace{\min_i \sum_{t=1}^{n} \ell_{ti}}_{\text{loss of best arm}}$$

Why this regret definition?

- The regret is with respect to the loss of the best arm

$$R_n = \underbrace{\mathbb{E}\left[\sum_{t=1}^{n} \ell_{tA_t}\right]}_{\text{learner's loss}} - \underbrace{\min_i \sum_{t=1}^{n} \ell_{ti}}_{\text{loss of best arm}}$$

- The following alternative objective is hopeless

$$R'_n = \underbrace{\mathbb{E}\left[\sum_{t=1}^{n} \ell_{tA_t}\right]}_{\text{learner's loss}} - \underbrace{\sum_{t=1}^{n} \min_i \ell_{ti}}_{\text{loss of best sequence}}$$

- Regret is at least cn for some $c > 1$.

$$\ell = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}$$

Tackling the adversarial bandit

- **Randomisation** is crucial in adversarial bandits
- Learner chooses distribution P_t over the K actions
- Samples $A_t \sim P_t$
- Observes the loss ℓ_{tA_t}

Tackling the adversarial bandit

- **Randomisation** is crucial in adversarial bandits
- Learner chooses distribution P_t over the K actions
- Samples $A_t \sim P_t$
- Observes the loss ℓ_{tA_t}
- Expected regret is

$$R_n = \max_i \mathbb{E}\left[\sum_{t=1}^n (\ell_{tA_t} - \ell_{ti})\right] = \max_{p \in \Delta^K} \mathbb{E}\left[\sum_{t=1}^n \langle P_t - p, \ell_t \rangle\right]$$

Tackling the adversarial bandit

- **Randomisation** is crucial in adversarial bandits
- Learner chooses distribution P_t over the K actions
- Samples $A_t \sim P_t$
- Observes the loss ℓ_{tA_t}
- Expected regret is

$$R_n = \max_i \mathbb{E}\left[\sum_{t=1}^n (\ell_{tA_t} - \ell_{ti})\right] = \max_{p \in \Delta^K} \mathbb{E}\left[\sum_{t=1}^n \langle P_t - p, \ell_t \rangle\right]$$

- How to choose P_t?

Tackling the adversarial bandit

- **Randomisation** is crucial in adversarial bandits
- Learner chooses distribution P_t over the K actions
- Samples $A_t \sim P_t$
- Observes the loss ℓ_{tA_t}
- Expected regret is

$$R_n = \max_i \mathbb{E}\left[\sum_{t=1}^n (\ell_{tA_t} - \ell_{ti})\right] = \max_{p \in \Delta^K} \mathbb{E}\left[\sum_{t=1}^n \langle P_t - p, \ell_t \rangle\right]$$

- How to choose P_t?
- Consider a simpler setting: choose the action A_t and the entire vector ℓ_t is observed (instead of ℓ_{tA_t})
- Online convex optimization with a linear loss

Online convex optimisation (linear losses)

- Domain of x $\mathcal{K} \subset \mathbb{R}^d$ is a convex set
- Adversary secretly chooses $\ell_1, \ldots, \ell_n \in \mathcal{K}^\circ = \{u : \sup_{x \in \mathcal{K}} |\langle x, u \rangle| \leq 1\}$ (polar set)
- At each time t, the learner chooses $x_t \in \mathcal{K}$
- Suffers loss $\langle x_t, \ell_t \rangle$

Online convex optimisation (linear losses)

- Domain of x $\mathcal{K} \subset \mathbb{R}^d$ is a convex set
- Adversary secretly chooses $\ell_1, \ldots, \ell_n \in \mathcal{K}^\circ = \{u : \sup_{x \in \mathcal{K}} |\langle x, u \rangle| \leq 1\}$ (polar set)
- At each time t, the learner chooses $x_t \in \mathcal{K}$
- Suffers loss $\langle x_t, \ell_t \rangle$
- The regret with respect to the best $x \in \mathcal{K}$ is

$$R_n(x) = \sum_{t=1}^{n} \langle x_t - x, \ell_t \rangle \,.$$

Online convex optimisation (linear losses)

- Domain of x $\mathcal{K} \subset \mathbb{R}^d$ is a convex set
- Adversary secretly chooses $\ell_1, \ldots, \ell_n \in \mathcal{K}^\circ = \{u : \sup_{x \in \mathcal{K}} |\langle x, u \rangle| \leq 1\}$ (polar set)
- At each time t, the learner chooses $x_t \in \mathcal{K}$
- Suffers loss $\langle x_t, \ell_t \rangle$
- The regret with respect to the best $x \in \mathcal{K}$ is

$$R_n(x) = \sum_{t=1}^n \langle x_t - x, \ell_t \rangle \,.$$

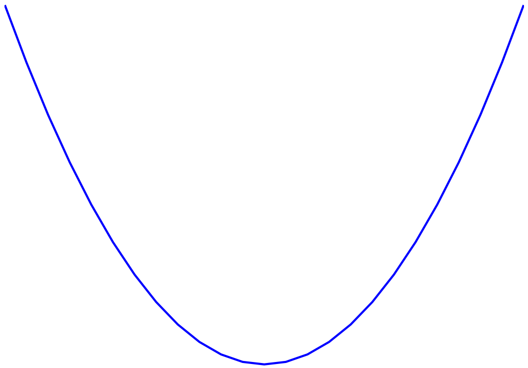
More general online convex optimization

- Learner chooses $x_t \in \mathcal{K}$
- Adversary chooses convex $f_t : \mathcal{K} \to \mathbb{R}$
- Suffer loss in round t is $f_t(x_t)$ and regret is

$$R_n(x) = \sum_{t=1}^n (f_t(x_t) - f_t(x))$$

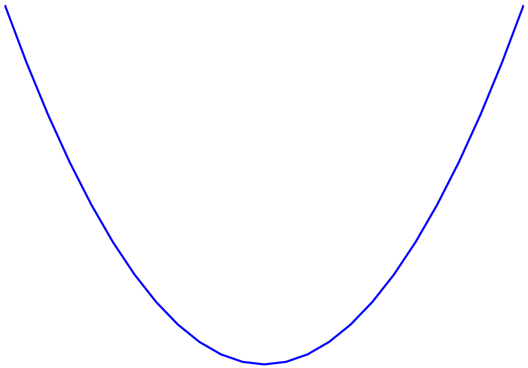
- linear is a special case with $f_t(x) = \langle x, \ell_t \rangle$

Why linear is enough?



- convex function
- The sum of convex functions is convex

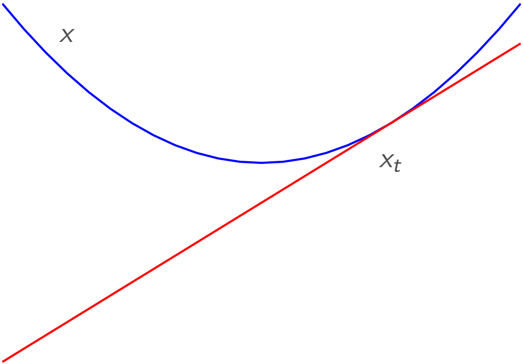
Why linear is enough?



- convex function
- The sum of convex functions is convex
- Strictly convex function has a unique minimizer

Linearisation of a convex function

$$f_t(x) \geq f_t(x_t) + \langle x - x_t, \nabla f_t(x_t) \rangle$$



Rearranging, $f_t(x_t) - f_t(x) \leq \langle x_t - x, \nabla f_t(x_t) \rangle$

Why linear is enough?

- Regret is bounded by

$$
\begin{aligned}
R_n(x) &= \sum_{t=1}^{n} (f_t(x_t) - f_t(x)) \\
&\leq \sum_{t=1}^{n} \langle x_t - x, \nabla f_t(x_t) \rangle
\end{aligned}
$$

- Reduction from nonlinear to linear
- Only uses first order information (the gradient)

Why linear is enough?

- Regret is bounded by

$$R_n(x) = \sum_{t=1}^{n} (f_t(x_t) - f_t(x)) \\ \le \sum_{t=1}^{n} \langle x_t - x, \nabla f_t(x_t) \rangle$$

- Reduction from nonlinear to linear
- Only uses first order information (the gradient)
- **Linear losses from now on** $\quad f_t(x) = \langle x, \ell_t \rangle$
- Think of $\ell_t = \nabla f_t(x_t)$ for a general convex loss function

Online convex optimisation (linear losses)

- Adversary secretly chooses $\ell_1, \ldots, \ell_n \in \mathcal{K}^\circ = \{u : \sup_{x \in \mathcal{K}} |\langle x, u \rangle| \leq 1\}$ (polar)
- Learner chooses $x_t \in \mathcal{K}$
- Suffers loss $\langle x_t, \ell_t \rangle$ and the regret with respect to $x \in \mathcal{K}$ is

$$R_n(x) = \sum_{t=1}^{n} \langle x_t - x, \ell_t \rangle \,.$$

Online convex optimisation (linear losses)

- Adversary secretly chooses $\ell_1, \ldots, \ell_n \in \mathcal{K}^\circ = \{u : \sup_{x \in \mathcal{K}} |\langle x, u \rangle| \leq 1\}$ (polar)
- Learner chooses $x_t \in \mathcal{K}$
- Suffers loss $\langle x_t, \ell_t \rangle$ and the regret with respect to $x \in \mathcal{K}$ is

$$R_n(x) = \sum_{t=1}^n \langle x_t - x, \ell_t \rangle \,.$$

- **How to choose $\mathbf{x}_t$?** Most simple idea 'follow-the-leader'

$$x_{t+1} = \operatorname{argmin}_{x \in \mathcal{K}} \sum_{s=1}^t \langle x, \ell_s \rangle \,.$$

- Fails miserably: $\mathcal{K} = [-1, 1]$, $\ell_1 = 1/2$, $\ell_2 = -1$, $\ell_3 = 1$, $\ldots$
- $x_1 = ?$, $x_2 = -1$ $(\operatorname{argmin}_{x \in \mathcal{K}} \langle x, \ell_1 \rangle)$, $x_3 = 1$ $(\operatorname{argmin}_{x \in \mathcal{K}} \langle x, \ell_1 + \ell_2 \rangle)$, $\ldots$
- $R_n(0) = \sum_{t=1}^n \langle x_t, \ell_t \rangle \approx n$.

Follow The regularized Leader (FTRL)

- **New idea** Add **regularization** to stabilize follow-the-leader
- Let F be a strictly convex function and $\eta > 0$ be the **learning rate** and

$$x_{t+1} = \mathrm{argmin}_{x \in \mathcal{K}} \left(F(x) + \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle \right)$$

Follow The regularized Leader (FTRL)

- **New idea** Add **regularization** to stabilize follow-the-leader
- Let F be a strictly convex function and $\eta > 0$ be the **learning rate** and

$$x_{t+1} = \operatorname{argmin}_{x \in \mathcal{K}} \left(F(x) + \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle \right)$$

- Different choices of F lead to different algorithms.
- One clean analysis.

Example – Gradient descent

- $\mathcal{K} = \mathbb{R}^d$ and $F(x) = \frac{1}{2}\|x\|_2^2$

$$x_{t+1} = \mathrm{argmin}_{x \in \mathcal{K}}\, \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle + \frac{1}{2}\|x\|_2^2$$

Example – Gradient descent

- $\mathcal{K} = \mathbb{R}^d$ and $F(x) = \frac{1}{2}\|x\|_2^2$

$$x_{t+1} = \operatorname{argmin}_{x \in \mathcal{K}} \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle + \frac{1}{2}\|x\|_2^2$$

- Differentiating,

$$0 = \eta \sum_{s=1}^{t} \ell_s + x$$

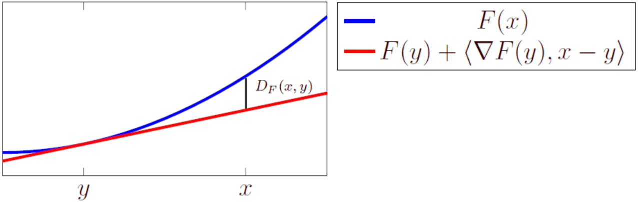
- $x_{t+1} = -\eta \sum_{s=1}^{t} \ell_s = x_t - \eta \ell_t$

A few tools

- Online convex optimization uses many tools from convex analysis
- Bregman divergence
- First-order optimality conditions
- Dual norms

Bregman divergence

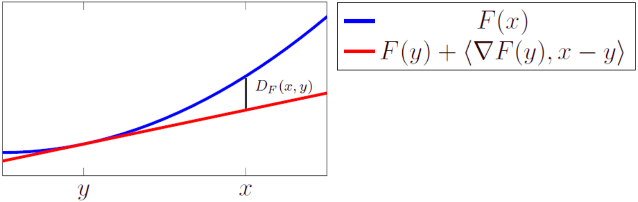
For convex F, $D_F(x, y) = F(x) - F(y) - \langle \nabla F(y), x - y \rangle$



- Bregman divergence is not a distance (may not be symmetric $D_F(x, y) \neq D_F(y, x)$, e.g., KL divergence), but still, $D_F(x, y) \geq 0$

Bregman divergence

For convex F, $D_F(x,y) = F(x) - F(y) - \langle \nabla F(y), x - y \rangle$



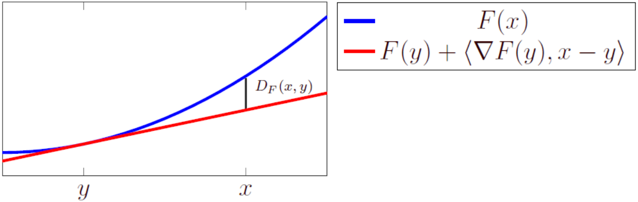
- Bregman divergence is not a distance (may not be symmetric $D_F(x,y) \neq D_F(y,x)$, e.g., KL divergence), but still, $D_F(x,y) \geq 0$

- By Taylor expansion, there exists a $z = \alpha x + (1-\alpha)y$ for $\alpha \in [0,1]$

$$D_F(x,y) = (x-y)^\top \nabla^2 F(z)(x-y) = \|x-y\|^2_{\nabla^2 F(z)}$$

Bregman divergence

For convex F, $D_F(x,y) = F(x) - F(y) - \langle \nabla F(y), x-y \rangle$



- Bregman divergence is not a distance (may not be symmetric $D_F(x,y) \neq D_F(y,x)$, e.g., KL divergence), but still, $D_F(x,y) \geq 0$

- By Taylor expansion, there exists a $z = \alpha x + (1-\alpha)y$ for $\alpha \in [0,1]$

$$D_F(x,y) = (x-y)^\top \nabla^2 F(z)(x-y) = \|x-y\|^2_{\nabla^2 F(z)}$$

- Key property: does not change under linear perturbation: For $\widetilde{F}(x) = F + \langle a, x \rangle$, $D_{\widetilde{F}}(x,y) = D_F(x,y)$

Examples

- **Quadratic** $F(x) = \frac{1}{2}\|x\|^2$

$$D_F(x, y) = \frac{1}{2}\|x - y\|_2^2$$

Examples

- **Quadratic** $F(x) = \frac{1}{2}\|x\|^2$

$$D_F(x,y) = \frac{1}{2}\|x-y\|_2^2$$

- **Neg-entropy** $F(x) = \sum_{i=1}^d x_i \log(x_i) - x_i$

$$D_F(x,y) = \sum_{i=1}^d x_i \log(\frac{x_i}{y_i}) + \sum_{i=1}^d (y_i - x_i)$$

When $x, y \in \Delta_d$, where $\Delta_d = \{x \in \mathbb{R}^d : x \geq 0, \|x\|_1 = 1\}$ (d-dimensional simplex, usually for modeling a discrete probability distribution):

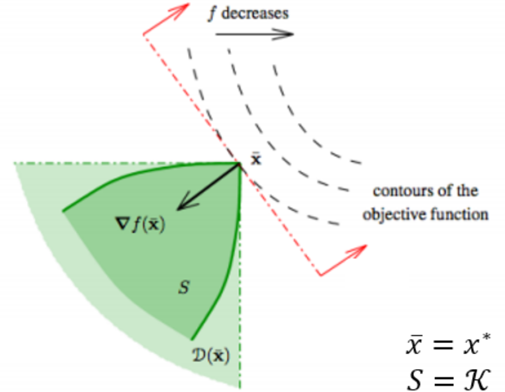
$$D_F(x,y) = \sum_{i=1}^d x_i \log(\frac{x_i}{y_i})$$

First order optimality condition

- Let $\mathcal{K}$ be convex, $f : \mathcal{K} \to \mathbb{R}$ convex, differentiable

$$x^* = \operatorname{argmin}_{x \in \mathcal{K}} f(x) \Leftrightarrow \langle \nabla f(x^*), x - x^* \rangle \geq 0 \;\forall x \in \mathcal{K}$$

- **Interpretation** f is increasing in direction $x - x^*$ for all $x \in \mathcal{K}$



$\bar{x} = x^*$
$S = \mathcal{K}$

Dual norm

Let $\|\cdot\|_t$ be a norm on $\mathbb{R}^d$, then its dual norm

$$\|z\|_{t^*} = \sup\{\langle z, x\rangle, \|x\|_t \leq 1\}.$$

Dual norm

Let $\|\cdot\|_t$ be a norm on $\mathbb{R}^d$, then its dual norm

$$\|z\|_{t^*} = \sup\{\langle z, x\rangle, \|x\|_t \leq 1\}.$$

The dual norm of $\|z\|_{t^*}$ will be $\|\cdot\|_t$.

Dual norm

Let $\| \cdot \|_t$ be a norm on $\mathbb{R}^d$, then its dual norm

$$\|z\|_{t^*} = \sup\{\langle z, x\rangle, \|x\|_t \leq 1\}.$$

The dual norm of $\|z\|_{t^*}$ will be $\| \cdot \|_t$.

- The dual norm of $\| \cdot \|_2$ is $\| \cdot \|_2$
- The dual norm of $\| \cdot \|_1$ is $\| \cdot \|_\infty$
- The dual norm of $\| \cdot \|_p$ is $\| \cdot \|_q$ (with $\frac{1}{p} + \frac{1}{q} = 1$).

Dual norm

Let $\|\cdot\|_t$ be a norm on $\mathbb{R}^d$, then its dual norm

$$\|z\|_{t^*} = \sup\{\langle z, x\rangle, \|x\|_t \leq 1\}.$$

The dual norm of $\|z\|_{t^*}$ will be $\|\cdot\|_t$.

- The dual norm of $\|\cdot\|_2$ is $\|\cdot\|_2$
- The dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$
- The dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ (with $\frac{1}{p} + \frac{1}{q} = 1$).
- Hölder's inequality: $\langle z, x\rangle \leq \|x\|_t \|z\|_{t^*}$.

Follow the regularized leader

- $x_{t+1} = \text{argmin}_{x \in \mathcal{K}} \left(F(x) + \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle \right)$
- Equivalent to

$$\begin{aligned} x_{t+1} &= \text{argmin}_{x \in \mathcal{K}} \left(\eta \langle x, \ell_t \rangle + D_F(x, x_t) \right) \\ &= \text{argmin}_{x \in \mathcal{K}} \left(\eta \langle x, \ell_t \rangle + F(x) - F(x_t) - \langle \nabla F(x_t), x - x_t \rangle \right) \end{aligned}$$

Follow the regularized leader

- $x_{t+1} = \mathrm{argmin}_{x \in \mathcal{K}} \left(F(x) + \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle \right)$
- Equivalent to

$$\begin{aligned} x_{t+1} &= \mathrm{argmin}_{x \in \mathcal{K}} \left(\eta \langle x, \ell_t \rangle + D_F(x, x_t) \right) \\ &= \mathrm{argmin}_{x \in \mathcal{K}} \left(\eta \langle x, \ell_t \rangle + F(x) - F(x_t) - \langle \nabla F(x_t), x - x_t \rangle \right) \end{aligned}$$

- Assuming the minimizer is achieved in the interior of K.
- The first optimization implies that $\nabla F(x_{t+1}) = -\eta \sum_{s=1}^{t} \ell_s$
- The second optimization implies that $\eta \ell_t + \nabla F(x_{t+1}) - \nabla F(x_t) = 0$ and thus
  $\nabla F(x_{t+1}) = -\eta \ell_t + \nabla F(x_t) = -\eta \sum_{s=1}^{t} \ell_s + \underbrace{\nabla F(x_1)}_{0} = -\eta \sum_{s=1}^{t} \ell_s.$

Regret Analysis: Follow the regularized leader

Theorem For any fixed action x, the regret of follow the regularized leader satisfies

$$
\begin{aligned}
R_n(x) &:= \sum_{t=1}^{n} \langle x_t - x, \ell_t \rangle \\
&\leq \frac{F(x) - F(x_1)}{\eta} + \sum_{t=1}^{n} \left(\langle x_t - x_{t+1}, \ell_t \rangle - \frac{1}{\eta} D_F(x_{t+1}, x_t) \right) \\
&\leq \frac{F(x) - F(x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_{t*}^2
\end{aligned}
$$

Let $z_t \in [x_t, x_{t+1}]$ be such that $D_F(x_{t+1}, x_t) = \frac{1}{2}\|x_t - x_{t+1}\|^2_{\nabla^2 F(z_t)}$

and $\|\cdot\|_t = \|\cdot\|_{\nabla^2 F(z_t)}$ and $\|\cdot\|_{t^*}$ is the dual norm of $\|\cdot\|_t$.

Regret Analysis: Follow the regularized leader

Theorem For any fixed action x, the regret of follow the regularized leader satisfies

$$
\begin{aligned}
R_n(x) &:= \sum_{t=1}^{n} \langle x_t - x, \ell_t \rangle \\
&\leq \frac{F(x) - F(x_1)}{\eta} + \sum_{t=1}^{n} \left(\langle x_t - x_{t+1}, \ell_t \rangle - \frac{1}{\eta} D_F(x_{t+1}, x_t) \right) \\
&\leq \frac{F(x) - F(x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_{t*}^2
\end{aligned}
$$

Let $z_t \in [x_t, x_{t+1}]$ be such that $D_F(x_{t+1}, x_t) = \frac{1}{2}\|x_t - x_{t+1}\|_{\nabla^2 F(z_t)}^2$

and $\|\cdot\|_t = \|\cdot\|_{\nabla^2 F(z_t)}$ and $\|\cdot\|_{t^*}$ is the dual norm of $\|\cdot\|_t$.

Choosing $\|\cdot\|_t$ such that $D_F(x_{t+1}, x_t) \geq \frac{1}{2}\|x_t - x_{t+1}\|_t^2$ is also valid.

Regret Analysis: Follow the regularized leader

Theorem For any fixed action x, the regret of follow the regularized leader satisfies

$$R_n(x) \leq \frac{F(x) - F(x_1)}{\eta} + \sum_{t=1}^{n} \left(\langle x_t - x_{t+1}, \ell_t \rangle - \frac{1}{\eta} D_F(x_{t+1}, x_t) \right)$$

$$\leq \frac{F(x) - F(x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_{t*}^2$$

$D_F(x_{t+1}, x_t) = \frac{1}{2}\|x_t - x_{t+1}\|^2_{\nabla^2 F(z_t)}$ and $\|\cdot\|_t = \|\cdot\|_{\nabla^2 F(z_t)}$

Proof of the second inequality:

$$\langle x_t - x_{t+1}, \ell_t \rangle - \frac{D_F(x_{t+1}, x_t)}{\eta} \leq \|\ell_t\|_{t*}\|x_t - x_{t+1}\|_t - \frac{D_F(x_{t+1}, x_t)}{\eta}$$

$$= \|\ell_t\|_{t*}\sqrt{2D_F(x_{t+1}, x_t)} - \frac{D_F(x_{t+1}, x_t)}{\eta} \leq \frac{\eta}{2}\|\ell_t\|_{t*}^2,$$

The last inequality is due to $ax - bx^2/2 \leq a^2/(2b)$ for any $b \geq 0$ with $a = \|\ell_t\|_{t*}$, $x = \sqrt{2D_F(x_{t+1}, x_t)}$ and $b = \frac{1}{\eta}$

FTRL analysis

- Proof of the first inequality

$$R_n(x) \leq \frac{F(x) - F(x_1)}{\eta} + \sum_{t=1}^{n} \left(\langle x_t - x_{t+1}, \ell_t \rangle - \frac{1}{\eta} D_F(x_{t+1}, x_t) \right)$$

- Rewriting the regret

$$\begin{aligned} R_n(x) &= \sum_{t=1}^{n} \langle x_t - x, \ell_t \rangle \\ &= \sum_{t=1}^{n} \langle x_t - x_{t+1}, \ell_t \rangle + \sum_{t=1}^{n} \langle x_{t+1} - x, \ell_t \rangle \end{aligned}$$

FTRL analysis

- Proof of the first inequality

$$R_n(x) \le \frac{F(x) - F(x_1)}{\eta} + \sum_{t=1}^{n} \left(\langle x_t - x_{t+1}, \ell_t \rangle - \frac{1}{\eta} D_F(x_{t+1}, x_t) \right)$$

- Rewriting the regret

$$\begin{aligned} R_n(x) &= \sum_{t=1}^{n} \langle x_t - x, \ell_t \rangle \\ &= \sum_{t=1}^{n} \langle x_t - x_{t+1}, \ell_t \rangle + \sum_{t=1}^{n} \langle x_{t+1} - x, \ell_t \rangle \end{aligned}$$

- Goal: show that

$$\sum_{t=1}^{n} \langle x_{t+1} - x, \ell_t \rangle \le \frac{F(x) - F(x_1)}{\eta} - \sum_{t=1}^{n} \frac{1}{\eta} D_F(x_{t+1}, x_t)$$

FTRL analysis

- Potential function: $\Phi_t(x) = \frac{F(x)}{\eta} + \sum_{s=1}^{t} \langle x, \ell_s \rangle$
- By FRTL: x_{t+1} minimizes Φ_t in $\mathcal{K}$
-

$$
\begin{aligned}
&\sum_{t=1}^{n} \langle x_{t+1} - x, \ell_t \rangle \\
&= \sum_{t=1}^{n} \langle x_{t+1}, \ell_t \rangle - \underbrace{(\sum_{t=1}^{n} \langle x, \ell_t \rangle + \frac{F(x)}{\eta})}_{\Phi_n(x)} + \frac{F(x)}{\eta} \\
&= \sum_{t=1}^{n} \underbrace{(\Phi_t(x_{t+1}) - \Phi_{t-1}(x_{t+1}))}_{\left(\frac{F(x_{t+1})}{\eta} + \sum_{s=1}^{t} \langle x_{t+1}, \ell_s \rangle\right) - \left(\frac{F(x_{t+1})}{\eta} + \sum_{s=1}^{t-1} \langle x_{t+1}, \ell_s \rangle\right)} - \Phi_n(x) + \frac{F(x)}{\eta},
\end{aligned}
$$

FTRL analysis

Potential function: $\Phi_t(x) = \frac{F(x)}{\eta} + \sum_{s=1}^{t} \langle x, \ell_s \rangle$ $(\Phi_0(x) = \frac{F(x)}{\eta})$

Then using: (1) $x_{t+1} = \mathrm{argmin}_x \Phi_t(x)$ and (2) $D_{\Phi_t}(\cdot,\cdot) = \frac{1}{\eta} D_F(\cdot,\cdot)$ (Bregman divergence keeps the same by adding linear functions)

$$\begin{aligned}
\sum_{t=1}^{n} \langle x_{t+1} - x, \ell_t \rangle &= \frac{F(x)}{\eta} + \sum_{t=1}^{n} (\Phi_t(x_{t+1}) - \Phi_{t-1}(x_{t+1})) - \Phi_n(x) \\
&= \frac{F(x)}{\eta} - \Phi_0(x_1) + \underbrace{\Phi_n(x_{n+1}) - \Phi_n(x)}_{\leq 0:\ x_{n+1} = \mathrm{argmin}_x \Phi_n(x)} + \sum_{t=0}^{n-1} (\Phi_t(x_{t+1}) - \Phi_t(x_{t+2})) \\
&\leq \frac{F(x) - F(x_1)}{\eta} + \sum_{t=0}^{n-1} (\Phi_t(x_{t+1}) - \Phi_t(x_{t+2})) \\
&= \frac{F(x) - F(x_1)}{\eta} - \sum_{t=0}^{n-1} \left(D_{\Phi_t}(x_{t+2}, x_{t+1}) + \underbrace{\langle \nabla \Phi_t(x_{t+1}), x_{t+2} - x_{t+1} \rangle}_{\geq 0} \right) \\
&\leq \frac{F(x) - F(x_1)}{\eta} - \frac{1}{\eta} \sum_{t=1}^{n} D_F(x_{t+1}, x_t),
\end{aligned}$$

where

$D_{\Phi_t}(x_{t+2}, x_{t+1}) = \Phi_t(x_{t+2}) - \Phi_t(x_{t+1}) - \langle \nabla \Phi_t(x_{t+1}), x_{t+2} - x_{t+1} \rangle.$

Final form of the regret

$$
\begin{aligned}
R_n(x) \leq & \frac{F(x) - F(x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_{t^*}^2 \\
\leq & \frac{\mathrm{diam}_F(\mathcal{K})}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_{t^*}^2 ,
\end{aligned}
$$

where $\mathrm{diam}_F(\mathcal{K}) := \max_{a,b \in \mathcal{K}} F(a) - F(b)$.

Final form of the regret

$$R_n(x) \leq \frac{F(x) - F(x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_{t^*}^2$$
$$\leq \frac{\operatorname{diam}_F(\mathcal{K})}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_{t^*}^2,$$

where $\operatorname{diam}_F(\mathcal{K}) := \max_{a,b \in \mathcal{K}} F(a) - F(b)$.

- Regret depends on distance from start to optimal
- Learning rate needs careful tuning

Application 1: Online gradient descent

Assume $\mathcal{K} = \{x : \|x\|_2 \leq 1\}$ and $\ell_t \in \mathcal{K}$ ($|\langle x, \ell_t \rangle| \leq 1$)

Choose $F(x) = \frac{1}{2}\|x\|_2^2$, $\mathrm{diam}_F(\mathcal{K}) := \max_{a,b\in\mathcal{K}} F(a) - F(b) = \frac{1}{2}$:

$$D_F(x, y) = \frac{1}{2}\|x - y\|_2^2$$

Application 1: Online gradient descent

Assume $\mathcal{K} = \{x : \|x\|_2 \leq 1\}$ and $\ell_t \in \mathcal{K}$ ($|\langle x, \ell_t \rangle| \leq 1$)

Choose $F(x) = \frac{1}{2}\|x\|_2^2$, $\mathrm{diam}_F(\mathcal{K}) := \max_{a,b \in \mathcal{K}} F(a) - F(b) = \frac{1}{2}$:

$$D_F(x, y) = \frac{1}{2}\|x - y\|_2^2$$

FTRL:

$$x_{t+1} = \mathrm{argmin}_{x \in \mathcal{K}}\, \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle + \frac{1}{2}\|x\|_2^2$$

Application 1: Online gradient descent

Assume $\mathcal{K} = \{x : \|x\|_2 \leq 1\}$ and $\ell_t \in \mathcal{K}$ ($|\langle x, \ell_t \rangle| \leq 1$)

Choose $F(x) = \frac{1}{2}\|x\|_2^2$, $\mathrm{diam}_F(\mathcal{K}) := \max_{a,b \in \mathcal{K}} F(a) - F(b) = \frac{1}{2}$:

$$D_F(x, y) = \frac{1}{2}\|x - y\|_2^2$$

FTRL:

$$x_{t+1} = \mathrm{argmin}_{x \in \mathcal{K}} \, \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle + \frac{1}{2}\|x\|_2^2$$

Then by choosing $\eta = \sqrt{1/n}$:

$$R_n(x) \leq \frac{\mathrm{diam}_F(\mathcal{K})}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_2^2 \leq \frac{1}{2\eta} + \frac{\eta n}{2} \leq \sqrt{n}$$

Application 2: Exponential weights

Assume $\mathcal{K} = \Delta_d := \{x \geq 0 : \|x\|_1 = 1\}$ and $\ell_t \in [0,1]^d$ for all t ($|\langle x, \ell_t \rangle| \leq 1$)

$$F(x) = \sum_{i=1}^{d} (x_i \log(x_i) - x_i)$$

$\mathrm{diam}_F(\mathcal{K}) := \max_{a,b \in \mathcal{K}} F(a) - F(b) = \log(d)$. This is because $\max_{x \in \mathcal{K}} F(x) = \log(d) - 1$ (achieving at $(1/d, \ldots, 1/d)$) and $\min_{x \in \mathcal{K}} F(x) = -1$ (achieving $(1, 0, \ldots, 0)$).

Application 2: Exponential weights

Assume $\mathcal{K} = \Delta_d := \{x \geq 0 : \|x\|_1 = 1\}$ and $\ell_t \in [0,1]^d$ for all t $(|\langle x, \ell_t \rangle| \leq 1)$

$$F(x) = \sum_{i=1}^{d} (x_i \log(x_i) - x_i)$$

$\mathrm{diam}_F(\mathcal{K}) := \max_{a,b \in \mathcal{K}} F(a) - F(b) = \log(d)$. This is because $\max_{x \in \mathcal{K}} F(x) = \log(d) - 1$ (achieving at $(1/d, \ldots, 1/d)$) and $\min_{x \in \mathcal{K}} F(x) = -1$ (achieving $(1, 0, \ldots, 0)$).

Bregman divergence

$$\begin{aligned} D_F(x, y) &= \sum_{i=1}^{d} x_i \log\left(\frac{x_i}{y_i}\right) && \text{(KL-divergence)} \\ &\geq \frac{1}{2}\|x - y\|_1^2 && \text{(Pinsker's inequality (exercise))} \end{aligned}$$

Application 2: Exponential weights

Assume $\mathcal{K} = \Delta_d := \{x \geq 0 : \|x\|_1 = 1\}$ and $\ell_t \in [0,1]^d$ for all t
FTRL:

$$x_{t+1} = \text{argmin}_{x \in \mathcal{K}} \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle + F(x)$$

Optimal action is a standard basis vector e_i, where i is the position that $i = \text{argmin}_{j=1,\dots,n} \left(\eta \sum_{s=1}^{t} \ell_{s,j}\right)$ (corresponding $F(x)$ is minimized since $F(e_i) = -1$).

Application 2: Exponential weights

Assume $\mathcal{K} = \Delta_d := \{x \geq 0 : \|x\|_1 = 1\}$ and $\ell_t \in [0,1]^d$ for all t
FTRL:

$$x_{t+1} = \mathrm{argmin}_{x \in \mathcal{K}}\, \eta \sum_{s=1}^{t} \langle x, \ell_s \rangle + F(x)$$

Optimal action is a standard basis vector e_i, where i is the position that $i = \mathrm{argmin}_{j=1,\dots,n} \left(\eta \sum_{s=1}^{t} \ell_{s,j}\right)$ (corresponding $F(x)$ is minimized since $F(e_i) = -1$).

$$\begin{aligned} R_n(x) &\leq \frac{\mathrm{diam}_F(\mathcal{K})}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_{t*}^2 \\ &\leq \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\ell_t\|_{\infty}^2 \\ &\leq \frac{\log(d)}{\eta} + \frac{\eta n}{2} \leq \sqrt{2n \log(d)} \end{aligned}$$

Our Goal: Adversarial bandits

- At the start of the game the **adversary** secretly chooses losses $\ell_1, \ldots, \ell_n$ with $\ell_t \in [0, 1]^K$
- In each round the learner chooses the arm $A_t \in \{1, \ldots, K\} \sim P_t$ (from some distribution P_t)
- Suffers and loss ℓ_{t,A_t} (only observe ℓ_{t,A_t})
- Regret is $R_n = \max_a \mathbb{E}\left[\sum_{t=1}^n \ell_{tA_t} - \ell_{ta}\right]$

Our Goal: Adversarial bandits

- At the start of the game the **adversary** secretly chooses losses $\ell_1, \ldots, \ell_n$ with $\ell_t \in [0,1]^K$
- In each round the learner chooses the arm $A_t \in \{1, \ldots, K\} \sim P_t$ (from some distribution P_t)
- Suffers and loss ℓ_{t,A_t} (only observe ℓ_{t,A_t})
- Regret is $R_n = \max_a \mathbb{E}\left[\sum_{t=1}^n \ell_{tA_t} - \ell_{ta}\right]$
- **Surprising result** there exists an algorithm such that $R_n \leq \sqrt{2nK\log(K)}$ for any adversary. How?

Our Goal: Adversarial bandits

- At the start of the game the **adversary** secretly chooses losses $\ell_1, \ldots, \ell_n$ with $\ell_t \in [0,1]^K$
- In each round the learner chooses the arm $A_t \in \{1, \ldots, K\} \sim P_t$ (from some distribution P_t)
- Suffers and loss ℓ_{t,A_t} (only observe ℓ_{t,A_t})
- Regret is $R_n = \max_a \mathbb{E}\left[\sum_{t=1}^{n} \ell_{tA_t} - \ell_{ta}\right]$
- **Surprising result** there exists an algorithm such that $R_n \leq \sqrt{2nK\log(K)}$ for any adversary. How?
- **Key idea**
    - Construct an estimator of the entire loss vector $\ell_t = (\ell_{t,1}, \ldots, \ell_{t,K})$
    - Apply the follow the regularized leader (FTRL) to the estimated loss $\hat{\ell}_t = (\hat{\ell}_{t,1}, \ldots, \hat{\ell}_{t,K})$

Importance-weighted estimators

At time t, our algorithm chooses the arm $A_t = i$ with probability P_{ti} (specify P_{ti} later). Define the estimator of $\ell_{t,i}$

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}(A_t = i)}{P_{ti}}$$

and $\hat{\ell}_t = (\hat{\ell}_{t,1}, \ldots \hat{\ell}_{t,K})$.

Unbiased estimator,

$$\begin{aligned} \mathbb{E}\left[\hat{\ell}_{t,i} \mid P_t\right] &= \frac{\ell_{t,i}}{P_{ti}}\mathbb{E}[\mathbb{1}(A_t = i) \mid P_t] = \frac{\ell_{t,i}}{P_{ti}}P_{ti} \\ &= \ell_{t,i} \end{aligned}$$

Second moment: $\mathbb{E}\left[\hat{\ell}^2_{t,i} \mid P_t\right] = \frac{\ell^2_{t,i}}{P_{ti}}$

Follow the regularized leader for bandits (EXP3)

- Estimate ℓ_t with unbiased **importance-weighted estimator** $\hat{\ell}_t$

$$\hat{\ell}_{t,i} = \frac{\mathbb{1}(A_t = i)\ell_{t,i}}{P_{ti}}$$

- Then the expected regret satisfies

$$\mathbb{E}[R_n] = \max_i \mathbb{E}\left[\sum_{t=1}^n \ell_{t,A_t} - \ell_{t,i}\right] = \max_i \mathbb{E}\left[\sum_{t=1}^n \langle P_t - e_i, \hat{\ell}_t \rangle\right]$$

This is because

- $\mathbb{E}(\langle P_t, \ell_t \rangle) = \mathbb{E}(\sum_{i=1}^K P_{ti}\ell_{t,i}) = \mathbb{E}(\sum_{i=1}^K \mathbb{1}(A_t = i)\ell_{t,i}) = \mathbb{E}(\ell_{t,A_t})$
- $\mathbb{E}(\langle e_i, \hat{\ell}_t \rangle) = \mathbb{E}(\hat{\ell}_{t,i}) = \ell_{t,i}$.

Follow the regularized leader for bandits (EXP3)

- Then the expected regret satisfies

$$\mathbb{E}[R_n] = \max_i \mathbb{E}\left[\sum_{t=1}^{n} \ell_{t,A_t} - \ell_{t,i}\right] = \max_i \mathbb{E}\left[\sum_{t=1}^{n} \langle P_t - e_i, \hat{\ell}_t \rangle\right]$$

- FTRL:

$$P_t = \operatorname{argmin}_{p \in \Delta_K} \frac{F(p)}{\eta} + \sum_{s=1}^{t-1} \langle p, \hat{\ell}_s \rangle$$

Follow the regularized leader for bandits (EXP3)

- Then the expected regret satisfies

$$\mathbb{E}[R_n] = \max_i \mathbb{E}\left[\sum_{t=1}^{n} \ell_{t,A_t} - \ell_{t,i}\right] = \max_i \mathbb{E}\left[\sum_{t=1}^{n} \langle P_t - e_i, \hat{\ell}_t \rangle\right]$$

- FTRL:

$$P_t = \text{argmin}_{p \in \Delta_K} \frac{F(p)}{\eta} + \sum_{s=1}^{t-1} \langle p, \hat{\ell}_s \rangle$$

- Since the domain is Δ_K, choose the **negentropy** F

$$F(p) = \sum_{i=1}^{K} p_i \log(p_i) - p_i$$

- Using Lagrange dual, the probability of choosing each arm i:

$$P_{ti} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,i}\right)}{\sum_{j=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,j}\right)}$$

Follow the regularized leader for bandits (EXP3 Algo)

- Using the FTRL regret bound:

$$\begin{aligned}
\mathbb{E}[R_n] &\leq \mathbb{E}\left[\frac{F(e_i) - F(P_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n} \|\hat{\ell}_t\|_{t*}^2\right] \\
&\leq \frac{\log(K)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n} \|\hat{\ell}_t\|_{t*}^2\right] \qquad (F(e_i) - F(P_1) \leq \log(K))
\end{aligned}$$

where i is the best arm.

Follow the regularized leader for bandits (EXP3 Algo)

- Using the FTRL regret bound:

$$\begin{aligned}\mathbb{E}[R_n] &\leq \mathbb{E}\left[\frac{F(e_i) - F(P_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n} \|\hat{\ell}_t\|_{t*}^2\right] \\ &\leq \frac{\log(K)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n} \|\hat{\ell}_t\|_{t*}^2\right] \qquad (F(e_i) - F(P_1) \leq \log(K))\end{aligned}$$

  where i is the best arm.

- How to bound $\|\hat{\ell}_t\|_{t*}^2$?

Follow the regularized leader for bandits

- How to bound $\|\hat{\ell}_t\|_{t*}^2$?

Follow the regularized leader for bandits

- How to bound $\|\hat{\ell}_t\|_{t*}^2$?
- Recall that Let $z_t \in [x_t, x_{t+1}]$ be such that $D_F(x_{t+1}, x_t) = \frac{1}{2}\|x_t - x_{t+1}\|^2_{\nabla^2 F(z_t)}$. And $\|\cdot\|_t = \|\cdot\|_{\nabla^2 F(z_t)}$ and $\|\cdot\|_{t^*}$ is the dual norm of $\|\cdot\|_t$.
- For $F(p) = \sum_{i=1}^K p_i \log(p_i) - p_i$,

$$\nabla^2 F(p) = \mathrm{diag}(1/p) \implies \|\hat{\ell}_t\|_{t*}^2 = \|\hat{\ell}_t\|^2_{\nabla^2 F(p)^{-1}} = \sum_{i=1}^K p_i \hat{\ell}_{t,i}^2,$$

for some $p \in [P_t, P_{t+1}]$

Follow the regularized leader for bandits

- How to bound $\|\hat{\ell}_t\|_{t*}^2$?

- $\|\hat{\ell}_t\|_{t*}^2 = \|\hat{\ell}_t\|_{\nabla^2 F(p)^{-1}}^2 = \sum_{i=1}^K p_i \hat{\ell}_{t,i}^2$ for some $p \in [P_t, P_{t+1}]$

Follow the regularized leader for bandits

- How to bound $\|\hat{\ell}_t\|_{t*}^2$?

- $\|\hat{\ell}_t\|_{t*}^2 = \|\hat{\ell}_t\|_{\nabla^2 F(p)^{-1}}^2 = \sum_{i=1}^K p_i \hat{\ell}_{t,i}^2$ for some $p \in [P_t, P_{t+1}]$

- $\hat{\ell}_{t,i} = \frac{\mathbb{1}(A_t = i)\ell_{t,i}}{P_{ti}}$ is non-negative and $\hat{\ell}_{t,j} = 0$ for $A_t \neq j$:

$$\|\hat{\ell}_t\|_{t*}^2 = p_{A_t} \hat{\ell}_{t,A_t}^2$$

Follow the regularized leader for bandits

- How to bound $\|\hat{\ell}_t\|_{t*}^2$?

- $\|\hat{\ell}_t\|_{t*}^2 = \|\hat{\ell}_t\|_{\nabla^2 F(p)^{-1}}^2 = \sum_{i=1}^K p_i \hat{\ell}_{t,i}^2$ for some $p \in [P_t, P_{t+1}]$

- $\hat{\ell}_{t,i} = \frac{\mathbb{1}(A_t=i)\ell_{t,i}}{P_{ti}}$ is non-negative and $\hat{\ell}_{t,j} = 0$ for $A_t \neq j$:

$$\|\hat{\ell}_t\|_{t*}^2 = p_{A_t} \hat{\ell}_{t,A_t}^2$$

- Further note that, $P_{t,A_t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,A_t}\right)}{\sum_{j=1}^K \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,j}\right)} := \frac{\alpha_{A_t}}{\sum_{j=1}^K \alpha_j}$ and

$$P_{t+1,A_t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,A_t}\right) \exp(-\eta \hat{\ell}_{t,A_t})}{\sum_{j=1}^K \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,j}\right) \exp(-\eta \hat{\ell}_{t,j})} = \frac{\alpha_{A_t}}{\alpha_{A_t} + \sum_{j \neq A_t} \alpha_j \exp(\eta \hat{\ell}_{t,A_t})}$$

Since $\exp(\eta \hat{\ell}_{t,A_t}) > 1$, $P_{t+1,A_t} \leq P_{t,A_t}$

$$\|\hat{\ell}_t\|_{t*}^2 = \sum_{j=1}^K p_j \hat{\ell}_{t,j}^2 \leq P_{tA_t} \hat{\ell}_{t,A_t}^2$$

Putting everything together: Regret bound for the follow the regularized leader for bandits

$$
\begin{aligned}
\mathbb{E}[R_n] &\leq \frac{\log(K)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n} \|\hat{\ell}_t\|_{t*}^2\right] \\
&\leq \frac{\log(K)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n} P_{tA_t}\hat{\ell}_{t,A_t}^2\right] \\
&= \frac{\log(K)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n} \frac{\ell_{t,A_t}^2}{P_{tA_t}}\right] \\
&\leq \frac{\log(K)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n} \frac{1}{P_{tA_t}}\right] && (\ell_t \in [0,1]^K) \\
&= \frac{\log(K)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{K} P_{ti} \cdot \frac{1}{P_{ti}}\right] \\
&= \frac{\log(K)}{\eta} + \frac{\eta n K}{2} \leq \sqrt{2nK\log(K)} && (\eta = \sqrt{2\log(K)/(nK)})
\end{aligned}
$$

Historical notes

- First paper on bandits is by Thompson (1933). He proposed an algorithm for two-armed Bernoulli bandits and hand-runs some simulations (Thompson sampling)
- Popularized enormously by Robbins (1952)
- Confidence bounds first used by Lai and Robbins (1985) to derive asymptotically optimal algorithm
- UCB by Katehakis and Robbins (1995) and Agrawal (1995). Finite-time analysis by Auer et al. (2002)
- Adversarial bandits: Auer et al. (1995)
- Minimax optimal algorithm by Audibert and Bubeck (2009)

Resources

- Online notes: http://banditalgs.com
- The book "Bandit Algorithms" by Tor Lattimore and Csaba Szepesvari
  https://tor-lattimore.com/downloads/book/book.pdf
- Book by Bubeck and Cesa-Bianchi (2012)
- Book by Cesa-Bianchi and Lugosi (2006)
- The Bayesian books by Gittins et al. (2011) and Berry and Fristedt (1985). Both worth reading.
- Notes by Aleksandrs Slivkins:
  http://slivkins.com/work/MAB-book.pdf

References I

Agrawal, R. (1995). Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078.

Audibert, J.-Y. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *Proceedings of Conference on Learning Theory (COLT)*, pages 217–226.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE.

Berry, D. and Fristedt, B. (1985). *Bandit problems : sequential allocation of experiments*. Chapman and Hall, London ; New York :.

Bubeck, S. and Cesa-Bianchi, N. (2012). *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning. Now Publishers Incorporated.

Bush, R. R. and Mosteller, F. (1953). A stochastic model with applications to learning. *The Annals of Mathematical Statistics*, pages 559–585.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

References II

Katehakis, M. N. and Robbins, H. (1995). Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.