On-line Learning

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

February 2023

A Game

Coin 1

$\mathbb{P}\{\text{heads}\} = p_1$

Coin 2

$\mathbb{P}\{\text{heads}\} = p_2$

Coin 3

$\mathbb{P}\{\text{heads}\} = p_3$

- p_1, p_2, and p_3 are **unknown**.
- You are given a total of 20 tosses.
- Maximise the total number of heads!

A Game

Coin 1

$\mathbb{P}\{\text{heads}\} = p_1$

Coin 2

$\mathbb{P}\{\text{heads}\} = p_2$

Coin 3

$\mathbb{P}\{\text{heads}\} = p_3$

- p_1, p_2, and p_3 are **unknown**.
- You are given a total of 20 tosses.
- Maximise the total number of heads!

Let's play!

A Game

Coin 1

$\mathbb{P}\{\text{heads}\} = p_1$

Coin 2

$\mathbb{P}\{\text{heads}\} = p_2$

Coin 3

$\mathbb{P}\{\text{heads}\} = p_3$

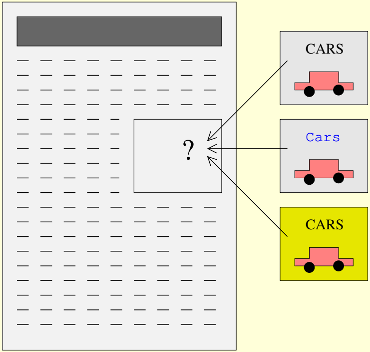
- p_1, p_2, and p_3 are **unknown**.
- You are given a total of 20 tosses.
- Maximise the total number of heads!

Let's play!

On-line learning: no "data" when we begin. Have to take actions to gather data.

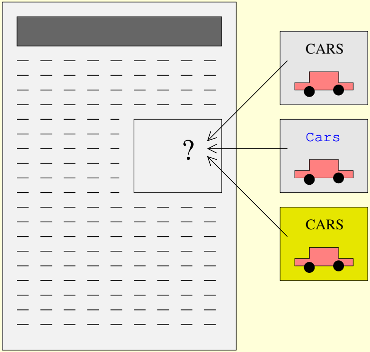
To Explore or to Exploit?

- On-line advertising: Template optimisation



To Explore or to Exploit?

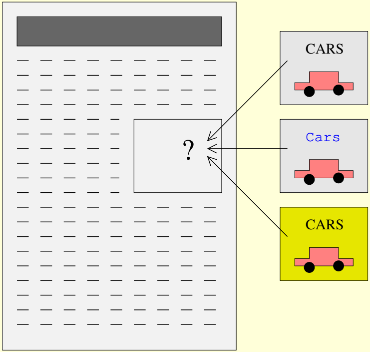
- On-line advertising: Template optimisation



- Clinical trials

To Explore or to Exploit?

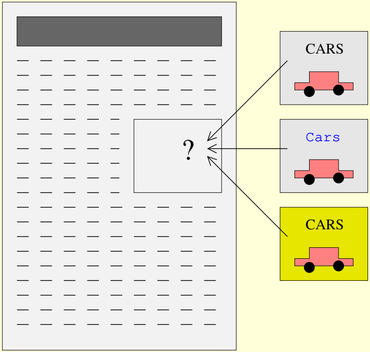
- On-line advertising: Template optimisation



- Clinical trials
- Packet routing in communication networks

To Explore or to Exploit?

- On-line advertising: Template optimisation

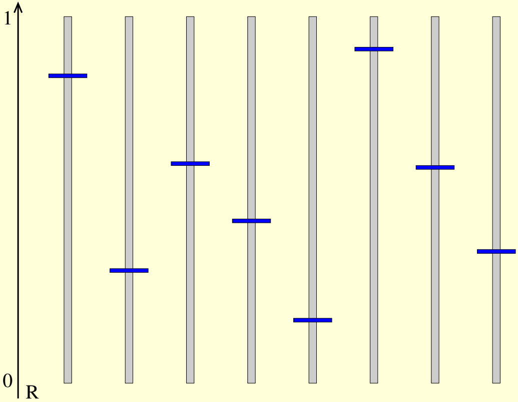


- Clinical trials
- Packet routing in communication networks
- Game playing and reinforcement learning

This lecture

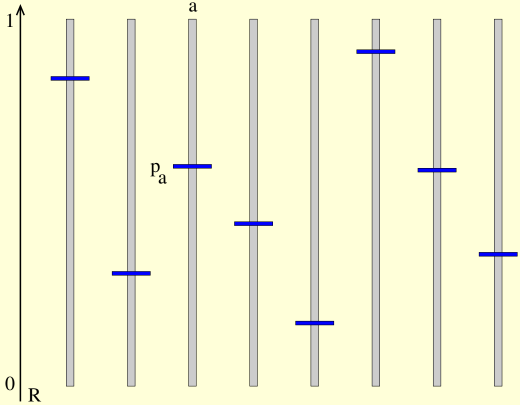
- Problem definition
- A natural algorithm
- Two improved algorithms
- Conclusion

Stochastic Multi-armed Bandits



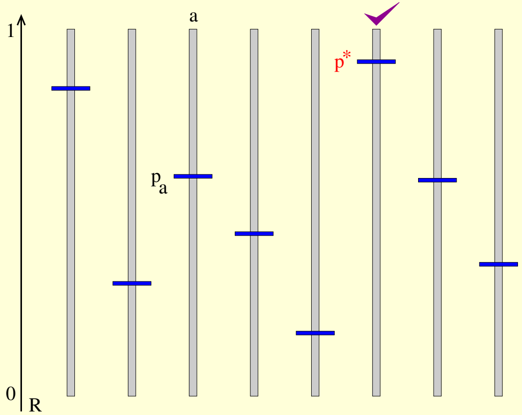
- *n* arms, each associated with a Bernoulli distribution (rewards are 0 or 1).

Stochastic Multi-armed Bandits



- n arms, each associated with a Bernoulli distribution (rewards are 0 or 1).
- Arm a has mean p_a.

Stochastic Multi-armed Bandits



- *n* arms, each associated with a Bernoulli distribution (rewards are 0 or 1).
- Arm a has mean p_a.
- Highest mean is p^*.

One-armed Bandits

[1]

1. https://pxhere.com/en/photo/942387.

Regret Minimisation

- What does an **algorithm** do?

Regret Minimisation

- What does an **algorithm** do?

  For $t = 1, 2, 3, \ldots, T$:

  - Given the history $a^1, r^1, a^2, r^2, a^3, r^3, \ldots, a^{t-1}, r^{t-1}$,
  - Pick an arm a^t to sample, and
  - Obtain a reward r^t drawn from the distribution corresponding to arm a^t.

Regret Minimisation

- What does an **algorithm** do?

  For $t = 1, 2, 3, \ldots, T$:
  - Given the history $a^1, r^1, a^2, r^2, a^3, r^3, \ldots, a^{t-1}, r^{t-1}$,
  - Pick an arm a^t to sample, and
  - Obtain a reward r^t drawn from the distribution corresponding to arm a^t.

- T is the total sampling budget, or the horizon.

Regret Minimisation

- What does an **algorithm** do?

  For $t = 1, 2, 3, \ldots, T$:

  - Given the history $a^1, r^1, a^2, r^2, a^3, r^3, \ldots, a^{t-1}, r^{t-1}$,
  - Pick an arm a^t to sample, and
  - Obtain a reward r^t drawn from the distribution corresponding to arm a^t.

- T is the total sampling budget, or the horizon.

- What is the maximum expected reward possible in T pulls?

Regret Minimisation

- What does an **algorithm** do?

  For $t = 1, 2, 3, \ldots, T$:

  - Given the history $a^1, r^1, a^2, r^2, a^3, r^3, \ldots, a^{t-1}, r^{t-1}$,
  - Pick an arm a^t to sample, and
  - Obtain a reward r^t drawn from the distribution corresponding to arm a^t.

- T is the total sampling budget, or the horizon.

- What is the maximum expected reward possible in T pulls? Tp^*.

Regret Minimisation

- What does an **algorithm** do?

  For $t = 1, 2, 3, \ldots, T$:

  - Given the history $a^1, r^1, a^2, r^2, a^3, r^3, \ldots, a^{t-1}, r^{t-1}$,
  - Pick an arm a^t to sample, and
  - Obtain a reward r^t drawn from the distribution corresponding to arm a^t.

- T is the total sampling budget, or the horizon.

- What is the maximum expected reward possible in T pulls? Tp^*.

- The actual expected reward for an algorithm is $\sum_{t=1}^{T} \mathbb{E}[r^t]$.

Regret Minimisation

- What does an **algorithm** do?

  For $t = 1, 2, 3, \ldots, T$:

  - Given the history $a^1, r^1, a^2, r^2, a^3, r^3, \ldots, a^{t-1}, r^{t-1}$,
  - Pick an arm a^t to sample, and
  - Obtain a reward r^t drawn from the distribution corresponding to arm a^t.

- T is the total sampling budget, or the horizon.

- What is the maximum expected reward possible in T pulls? Tp^*.

- The actual expected reward for an algorithm is $\sum_{t=1}^{T} \mathbb{E}[r^t]$.

- The regret of the algorithm is the difference

$$R_T = Tp^* - \sum_{t=1}^{T} \mathbb{E}[r^t].$$

We desire an algorithm that minimises regret!

Regret Minimisation

- What does an **algorithm** do?

  For $t = 1, 2, 3, \ldots, T$:
  - Given the history $a^1, r^1, a^2, r^2, a^3, r^3, \ldots, a^{t-1}, r^{t-1}$,
  - Pick an arm a^t to sample, and
  - Obtain a reward r^t drawn from the distribution corresponding to arm a^t.

- T is the total sampling budget, or the horizon.

- What is the maximum expected reward possible in T pulls? Tp^*.

- The actual expected reward for an algorithm is $\sum_{t=1}^{T} \mathbb{E}[r^t]$.

- The regret of the algorithm is the difference

$$R_T = Tp^* - \sum_{t=1}^{T} \mathbb{E}[r^t].$$

We desire an algorithm that minimises regret! Can you think of one?

This Lecture

- Problem definition
- A natural algorithm
- Two improved algorithms
- Conclusion

ϵ-Greedy Strategies

- ϵG1 (parameter $\epsilon \in [0, 1]$ controls the amount of exploration)
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - At $t = \lfloor \epsilon T \rfloor$, identify a^{best}, an arm with the highest empirical mean.
  - If $t > \epsilon T$, sample a^{best}.

ϵ-Greedy Strategies

- ϵG1 (parameter $\epsilon \in [0, 1]$ controls the amount of exploration)
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - At $t = \lfloor \epsilon T \rfloor$, identify a^{best}, an arm with the highest empirical mean.
  - If $t > \epsilon T$, sample a^{best}.

- Test instance I_1: $n = 20$; means = $0.01, 0.02, 0.03, \ldots, 0.2$; $T = 100,000$.

ϵ-Greedy Strategies

- ϵG1 (parameter $\epsilon \in [0, 1]$ controls the amount of exploration)
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - At $t = \lfloor \epsilon T \rfloor$, identify a^{best}, an arm with the highest empirical mean.
  - If $t > \epsilon T$, sample a^{best}.

- Test instance I_1: $n = 20$; means = $0.01, 0.02, 0.03, \ldots, 0.2$; $T = 100,000$.

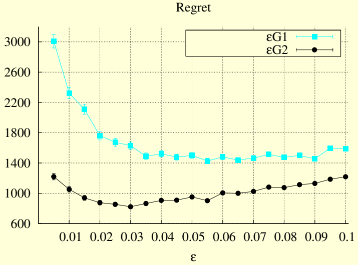
ϵ-Greedy Strategies

- ϵG1 (parameter $\epsilon \in [0, 1]$ controls the amount of exploration)
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - At $t = \lfloor \epsilon T \rfloor$, identify a^{best}, an arm with the highest empirical mean.
  - If $t > \epsilon T$, sample a^{best}.
- ϵG2
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - If $t > \epsilon T$, sample an arm with the highest empirical mean.

- Test instance I_1: $n = 20$; means = $0.01, 0.02, 0.03, \ldots, 0.2$; $T = 100,000$.

ϵ-Greedy Strategies

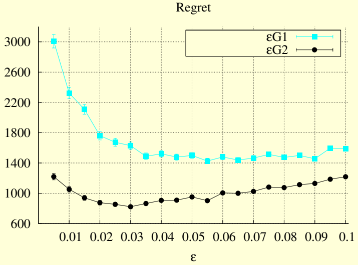
- ϵG1 (parameter $\epsilon \in [0, 1]$ controls the amount of exploration)
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - At $t = \lfloor \epsilon T \rfloor$, identify a^{best}, an arm with the highest empirical mean.
  - If $t > \epsilon T$, sample a^{best}.
- ϵG2
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - If $t > \epsilon T$, sample an arm with the highest empirical mean.



- Test instance I_1: $n = 20$; means = $0.01, 0.02, 0.03, \ldots, 0.2$; $T = 100,000$.

ϵ-Greedy Strategies

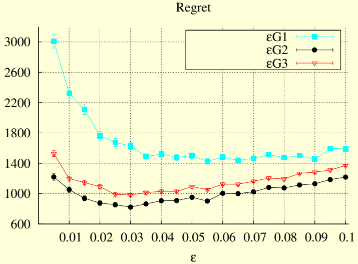
- ϵG1 (parameter $\epsilon \in [0, 1]$ controls the amount of exploration)
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - At $t = \lfloor \epsilon T \rfloor$, identify a^{best}, an arm with the highest empirical mean.
  - If $t > \epsilon T$, sample a^{best}.
- ϵG2
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - If $t > \epsilon T$, sample an arm with the highest empirical mean.
- ϵG3
  - With probability ϵ, sample an arm uniformly at random; with probability $1 - \epsilon$, sample an arm with the highest empirical mean.



- Test instance I_1: $n = 20$; means = $0.01, 0.02, 0.03, \ldots, 0.2$; $T = 100,000$.

ϵ-Greedy Strategies

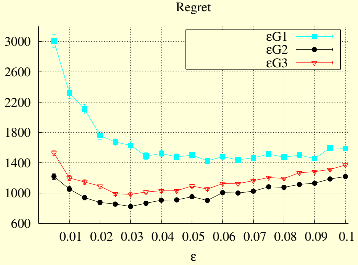
- ϵG1 (parameter $\epsilon \in [0, 1]$ controls the amount of exploration)
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - At $t = \lfloor \epsilon T \rfloor$, identify a^{best}, an arm with the highest empirical mean.
  - If $t > \epsilon T$, sample a^{best}.
- ϵG2
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - If $t > \epsilon T$, sample an arm with the highest empirical mean.
- ϵG3
  - With probability ϵ, sample an arm uniformly at random; with probability $1 - \epsilon$, sample an arm with the highest empirical mean.



- Test instance I_1: $n = 20$; means = $0.01, 0.02, 0.03, \ldots, 0.2$; $T = 100,000$.

ϵ-Greedy Strategies

- ϵG1 (parameter $\epsilon \in [0, 1]$ controls the amount of exploration)
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - At $t = \lfloor \epsilon T \rfloor$, identify a^{best}, an arm with the highest empirical mean.
  - If $t > \epsilon T$, sample a^{best}.
- ϵG2
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - If $t > \epsilon T$, sample an arm with the highest empirical mean.
- ϵG3
  - With probability ϵ, sample an arm uniformly at random; with probability $1 - \epsilon$, sample an arm with the highest empirical mean.



- Test instance I_1: $n = 20$; means = $0.01, 0.02, 0.03, \ldots, 0.2$; $T = 100,000$.

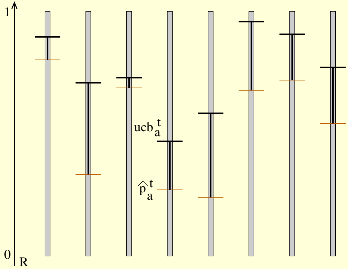
ϵG2 with $\epsilon = 0.03$ denoted ϵG^*. Regret of 822 ± 24 over a horizon of $100,000$.

This Lecture

- Problem definition
- A natural algorithm
- Two improved algorithms
- Conclusion

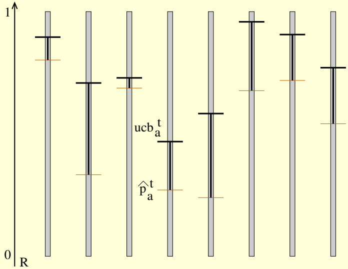
Upper Confidence Bounds

- UCB (Auer et al., 2002)
  - At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
  - $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
  - u_a^t the number of times a has been sampled at time t.



Upper Confidence Bounds

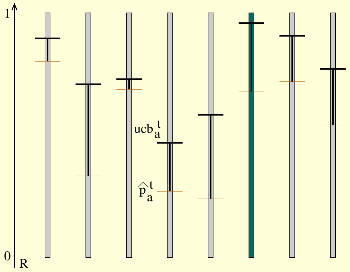
- UCB (Auer et al., 2002)
  - At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
  - $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
  - u_a^t the number of times a has been sampled at time t.



  - Sample an arm a for which ucb_a^t is maximal.

Upper Confidence Bounds

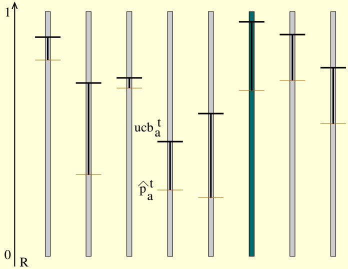
- UCB (Auer et al., 2002)
  - At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
  - $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
  - u_a^t the number of times a has been sampled at time t.



  - Sample an arm a for which ucb_a^t is maximal.

Upper Confidence Bounds

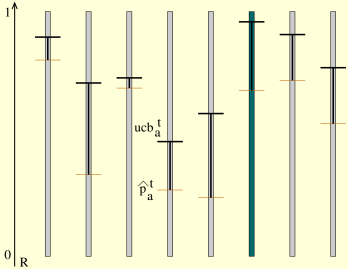
- UCB (Auer et al., 2002)
  - At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
  - $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
  - u_a^t the number of times a has been sampled at time t.



  - Sample an arm a for which ucb_a^t is maximal.
- Achieves regret of $O(\log(T))$: optimal dependence on T.

Upper Confidence Bounds

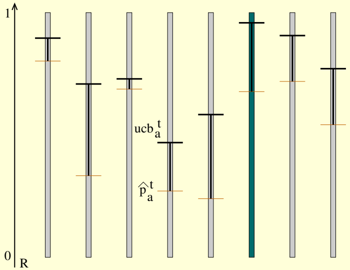
- UCB (Auer et al., 2002)
  - At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
  - $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
  - u_a^t the number of times a has been sampled at time t.



  - Sample an arm a for which ucb_a^t is maximal.
- Achieves regret of $O(\log(T))$: optimal dependence on T.
- KL-UCB (Garivier and Cappé, 2011) improves the constant in the $O()$.

Upper Confidence Bounds

- UCB (Auer et al., 2002)
  - At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
  - $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
  - u_a^t the number of times a has been sampled at time t.



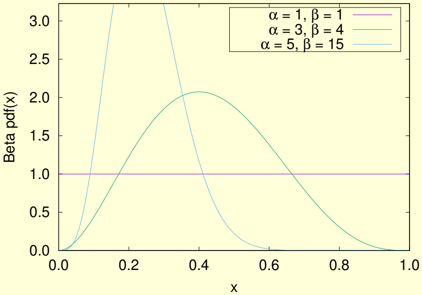
  - Sample an arm a for which ucb_a^t is maximal.
- Achieves regret of $O(\log(T))$: optimal dependence on T.
- KL-UCB (Garivier and Cappé, 2011) improves the constant in the $O()$.

Regret on instance I_1 (with $T = 100,000$)–UCB: 1168 ± 16; KL-UCB: 738 ± 18.

Before Moving on ... The Beta Distribution

- Beta(α, β) defined on $[0, 1]$.
  Two parameters: α and β.

$$\text{Mean} = \frac{\alpha}{\alpha + \beta}; \quad \text{Variance} = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

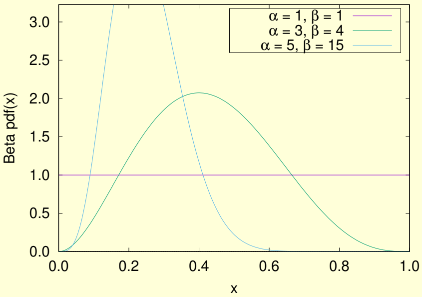


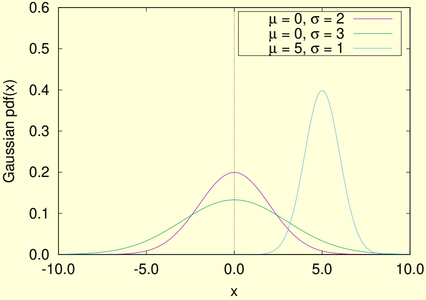
Plots obtained by adapting gnuplot script `http://gnuplot.sourceforge.net/demo/prob.5.gnu`.

Before Moving on … The Beta Distribution

- Beta(α, β) defined on [0, 1].
  Two parameters: α and β.

$$\text{Mean} = \frac{\alpha}{\alpha+\beta}; \quad \text{Variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$





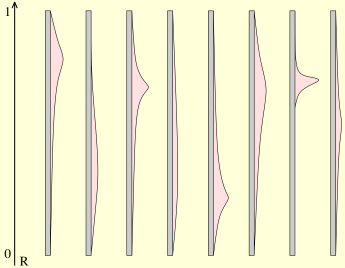
Plots obtained by adapting gnuplot script `http://gnuplot.sourceforge.net/demo/prob.5.gnu`.

Thompson Sampling

- Thompson (Thompson, 1933)
  - At time t, let arm *a* have s_a^t successes (ones/heads) and f_a^t failures (zeroes/tails).

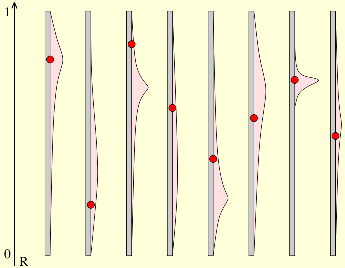
Thompson Sampling

- Thompson (Thompson, 1933)
  - At time t, let arm a have s_a^t successes (ones/heads) and f_a^t failures (zeroes/tails).
  - $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about the true mean of arm a.
  - Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$.



Thompson Sampling

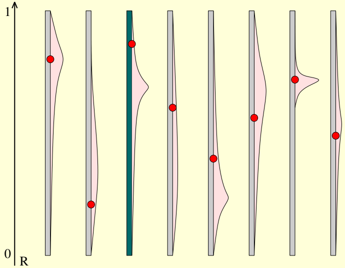
- Thompson (Thompson, 1933)
  - At time t, let arm a have s_a^t successes (ones/heads) and f_a^t failures (zeroes/tails).
  - $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about the true mean of arm a.
  - Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$.



  - Computational step: For every arm a, draw a sample $x_a^t \sim Beta(s_a^t + 1, f_a^t + 1)$.
  - Sampling step: Sample an arm a for which x_a^t is maximal.

Thompson Sampling

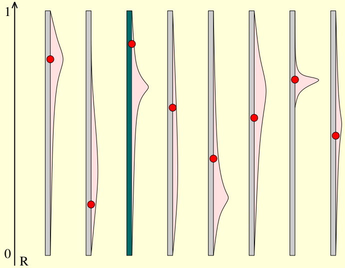
- Thompson (Thompson, 1933)
  - At time t, let arm a have s_a^t successes (ones/heads) and f_a^t failures (zeroes/tails).
  - $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about the true mean of arm a.
  - Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$.



  - Computational step: For every arm a, draw a sample $x_a^t \sim Beta(s_a^t + 1, f_a^t + 1)$.
  - Sampling step: Sample an arm a for which x_a^t is maximal.

Thompson Sampling

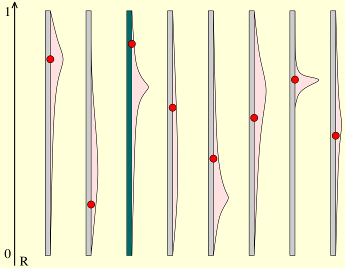
- Thompson (Thompson, 1933)
  - At time t, let arm a have s_a^t successes (ones/heads) and f_a^t failures (zeroes/tails).
  - $Beta(s_a^t + 1, f_a^t + 1)$ represents a “belief” about the true mean of arm a.
  - Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$.



  - Computational step: For every arm a, draw a sample $x_a^t \sim Beta(s_a^t + 1, f_a^t + 1)$.
  - Sampling step: Sample an arm a for which x_a^t is maximal.
- Achieves optimal regret (Kaufmann et al., 2012); is excellent in practice (Chapelle and Li, 2011).

Thompson Sampling

- Thompson (Thompson, 1933)
  - At time t, let arm *a* have s_a^t successes (ones/heads) and f_a^t failures (zeroes/tails).
  - $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about the true mean of arm *a*.
  - Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$.



  - Computational step: For every arm *a*, draw a sample $x_a^t \sim Beta(s_a^t + 1, f_a^t + 1)$.
  - Sampling step: Sample an arm *a* for which x_a^t is maximal.
- Achieves optimal regret (Kaufmann et al., 2012); is excellent in practice (Chapelle and Li, 2011).

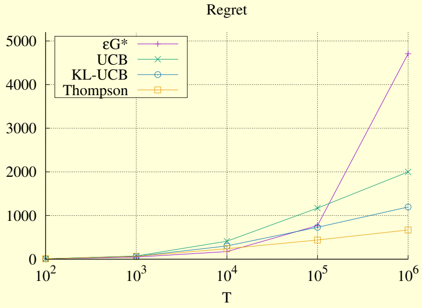
> On instance I_1 (with $T = 100,000$), regret is 463 ± 18.

Consolidated Results on Instance I_1

Method	Regret at $T = 100,000$
ϵG^*	822 ± 24
UCB	1168 ± 16
KL-UCB	738 ± 16
Thompson	463 ± 18

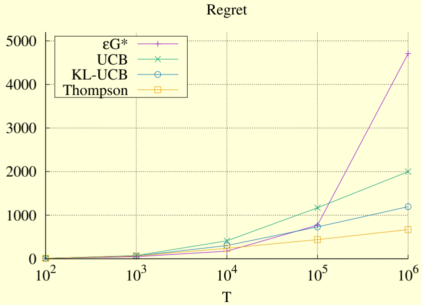
Consolidated Results on Instance I_1

Method	Regret at T $= 100,000$
ϵG*	822 ± 24
UCB	1168 ± 16
KL-UCB	738 ± 16
Thompson	463 ± 18



Consolidated Results on Instance I_1

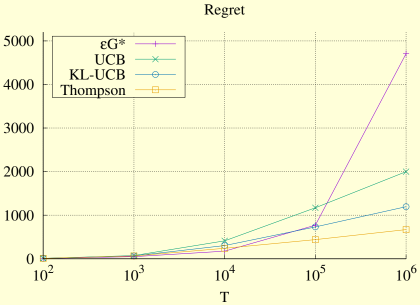
Method	Regret at $T = 100,000$
ϵG*	822 ± 24
UCB	1168 ± 16
KL-UCB	738 ± 16
Thompson	463 ± 18



Use Thompson Sampling!

Consolidated Results on Instance I_1

Method	Regret at $T = 100,000$
ϵG^*	822 ± 24
UCB	1168 ± 16
KL-UCB	738 ± 16
Thompson	463 ± 18



Use Thompson Sampling!

Principle: "Optimism in the face of uncertainty."

This Lecture

- Problem definition
- A natural algorithm
- Two improved algorithms
- Conclusion

Discussion

- **Challenges and extensions**

Discussion

- **Challenges and extensions**
  - Set of arms can change over time.

Discussion

- **Challenges and extensions**
  - Set of arms can change over time.
  - On-line updates not feasible; batch updating needed.

Discussion

- **Challenges and extensions**
  - Set of arms can change over time.
  - On-line updates not feasible; batch updating needed.
  - Rewards are delayed.

Discussion

- **Challenges and extensions**
  - Set of arms can change over time.
  - On-line updates not feasible; batch updating needed.
  - Rewards are delayed.
  - Arms might be *dependent*; “context” can be modeled.

Discussion

- **Challenges and extensions**
  - Set of arms can change over time.
  - On-line updates not feasible; batch updating needed.
  - Rewards are delayed.
  - Arms might be *dependent*; “context” can be modeled.
  - Nonstationary rewards (changing over time); adversarial modeling possible.

Discussion

- **Challenges and extensions**
  - Set of arms can change over time.
  - On-line updates not feasible; batch updating needed.
  - Rewards are delayed.
  - Arms might be *dependent*; “context” can be modeled.
  - Nonstationary rewards (changing over time); adversarial modeling possible.
- **Summary**
  - Adaptive sampling of options, based on stochastic feedback, to maximise total reward.
  - Well-studied problem with long history.
  - Thompson Sampling is an essentially optimal algorithm.
  - Modeling assumptions typically violated only slightly in practice.

References

- Chapter 2, **Reinforcement Learning: An Introduction**, Richard S. Sutton and Andrew G. Barto, 2020. Available on-line at `http://www.incompleteideas.net/book/RLbook2020.pdf`.
- **An Empirical Evaluation of Thompson Sampling**. Olivier Chapelle and Lihong Li, Neural Information Processing Systems 2011. Available on-line at `https://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling.pdf`.