# RLearning: Short guides to reinforcement learning

## Unit 1-4: Thompson Sampling

Davud Rostam-Afschar (Uni Mannheim)

# How to update your priors about rewards?

## Thompson Sampling

- Notation:
  - $r_t^a = r_t | A_t = a$ random variable for $a$'s rewards
  - $R(a) = q(a) = \mathbb{E}[r_t^a]$ unknown average reward
- Idea:
  - Sample several potential average rewards:
    $R_1(a), \dots, R_d(a) \sim \mathbb{P}(R(a) \mid r_1^a, \dots, r_t^a)$ for each $a$
  - Sample empirical average

$$\hat{R}(a) = \frac{1}{d} \sum_{i=1}^{d} R_i(a)$$

  - Execute $\underset{a}{\text{argmax}}\, \hat{R}(a)$
  - Coin example
  - $\mathbb{P}(R(a) \mid r_1^a, \dots, r_t^a) = \text{Beta}(\theta_a; \alpha_a, \beta_a)$
    where $\alpha_a - 1 = \#\text{heads}$ and $\beta_a - 1 = \#$ tails

# Bayesian Learning

- Notation:
  - $\mathbb{P}\left(r^a; \theta\right)$: unknown distribution (parameterized by $\theta$)
- Idea:
  - Express uncertainty about $\theta$ by a prior $\mathbb{P}(\theta)$
  - Compute posterior $\mathbb{P}\left(\theta \mid r_1^a, r_2^a, \ldots, r_t^a\right)$ based on
  - Samples $r_1^a, r_2^a, \ldots, r_t^a$ observed for $a$ so far
- Bayes theorem:

$$\mathbb{P}\left(\theta \mid r_1^a, r_2^a, \ldots, r_t^a\right) \propto \mathbb{P}\left(\theta\right) \mathbb{P}\left(r_1^a, r_2^a, \ldots, r_t^a \mid \theta\right)$$

# Distributional Information

- Posterior over $\theta$ allows us to estimate
  - Distribution over next reward $r^a$

$$\mathbb{P}\left(r^a \mid r_1^a, r_2^a, \ldots, r_t^a\right) = \int_\theta \mathbb{P}\left(r^a; \theta\right) \mathbb{P}\left(\theta \mid r_1^a, r_2^a, \ldots, r_t^a\right) d\theta$$

  - Distribution over $R(a)$ when $\theta$ includes the mean

$$\mathbb{P}\left(R(a) \mid r_1^a, r_2^a, \ldots, r_t^a\right) = \mathbb{P}\left(\theta \mid r_1^a, r_2^a, \ldots, r_t^a\right) \text{ if } \theta = R(a)$$

- To guide exploration:
  - UCB: $\mathbb{P}\left(R(a) > \text{bound}\left(r_1^a, r_2^a, \ldots, r_t^a\right)\right) \geq p$
  - Bayesian techniques: $\mathbb{P}\left(R(a) \mid r_1^a, r_2^a, \ldots, r_t^a\right)$

# Coin Example

- Consider two biased coins $C_1$ and $C_2$

$$R(C_1) = \mathbb{P}(C_1 = \text{ head})$$
$$R(C_2) = \mathbb{P}(C_2 = \text{ head})$$

- Problem:
  - Maximize # of heads in $d$ flips
  - Which coin should we choose for each flip?

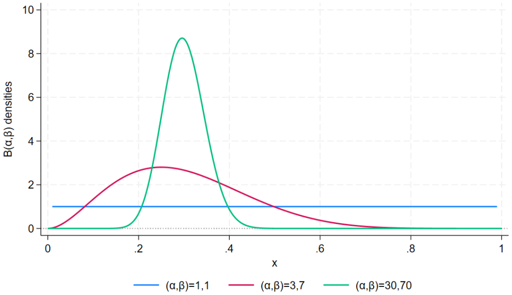
- $r^{c_1}, r^{c_2}$ are Bernoulli variables with domain $\{0, 1\}$
- Bernoulli dist. are parameterized by their mean

$$\text{i.e. } \mathbb{P}\left(r^{C_1}; \theta_1\right) = \theta_1 = R\left(C_1\right)$$
$$\mathbb{P}\left(r^{C_2}; \theta_2\right) = \theta_2 = R\left(C_2\right)$$

# Beta Distribution

- Let the prior $\mathbb{P}(\theta)$ be a Beta distribution
  $\text{Beta}(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

- $\alpha - 1$: # of heads
- $\beta - 1$: # of tails
- $\mathbb{E}[\theta] = \alpha/(\alpha+\beta)$



# Belief Update

- Prior: $\mathbb{P}(\theta) = \text{Beta}(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- Posterior after coin flip:

$$\begin{aligned}
\mathbb{P}(\theta \mid \text{head}) &\propto \quad \mathbb{P}(\theta) \qquad\qquad \mathbb{P}(\text{head} \mid \theta) \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \theta \\
&= \theta^{(\alpha+1)-1}(1-\theta)^{\beta-1} \\
&\propto \text{Beta}(\theta; \alpha+1, \beta) \\
\mathbb{P}(\theta \mid \text{tail}) &\propto \quad \mathbb{P}(\theta) \qquad\qquad \mathbb{P}(\text{tail} \mid \theta) \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad (1-\theta) \\
&= \theta^{\alpha-1}(1-\theta)^{(\beta+1)-1} \\
&\propto \text{Beta}(\theta; \alpha, \beta+1)
\end{aligned}$$

# Thompson Sampling Algorithm: Bernoulli Rewards

ThompsonSampling ($T$)

$V \leftarrow 0$

For $t = 1$ to $T$

Sample $R_1(a), \ldots, R_d(a) \sim \mathbb{P}(R(a)) \; \forall a$

$\hat{R}(a) \leftarrow \frac{1}{d} \sum_{i=1}^{d} R_i(a) \; \forall a$

$a^* \leftarrow \underset{\text{argmax}}{a} \hat{R}(a)$

Execute $a^*$ and receive $r$

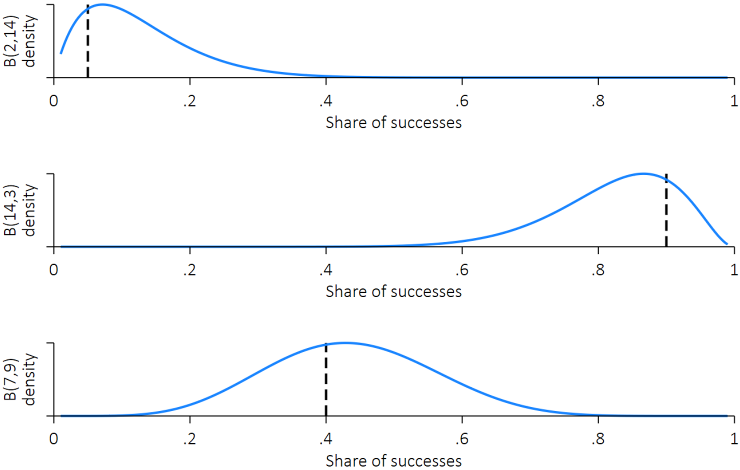
$V \leftarrow V + r$

Update $\mathbb{P}\left(R\left(a^*\right)\right)$ based on $r$

Return $V$

# Exploration vs Exploitation

# ?? Sampling




- Beta-Bernoulli Thompson sampling
- Models uncertainty about the shape of the distribution and the expected outcome $R$ explicitly

**Click to watch!**

# Comparison

**Thompson Sampling**

- Samples
  $r_i^a \sim \mathbb{P}(r^a; \theta)$
  $R_i(a) \sim \mathbb{P}(R_i(a) \mid r_1^a \dots r_t^a)$
- Empirical mean
  $\hat{R}(a) = \frac{1}{d} \sum_{i=1}^{d} R_i(a)$
- Action Selection
  $a^* = \underset{a}{\operatorname{argmax}}\ \hat{R}(a)$
- Some exploration

**Greedy Strategy**

- Samples
  $r_i^a \sim \mathbb{P}(r^a; \theta)$
- Empirical mean
  $\tilde{R}(a) = \frac{1}{t} \sum_{i=1}^{t} r_i^a$
- Action Selection
  $a^* = \underset{a}{\operatorname{argmax}}\ \tilde{R}(a)$
- No exploration

(**?**)

# Sample Size

- In Thompson sampling, amount of data $t$ and sample size $d$ regulate amount of exploration
- As $t$ and $d$ increase, $\hat{R}(a)$ becomes less stochastic, which reduces exploration
  - As $t \uparrow, \mathbb{P}\left(R(a) \mid r_1^a, \ldots, r_t^a\right)$ becomes more peaked
  - As $d \uparrow, \hat{R}(a)$ approaches $\mathbb{E}\left[R(a) \mid r_1^a, \ldots, r_t^a\right]$
- The stochasticity of $\hat{R}(a)$ ensures that all actions are chosen with some probability

# Analysis

- Thompson sampling converges to best arm
- Theory:
  - Expected cumulative regret: $\mathcal{O}(\log T)$
  - On par with UCB and $\varepsilon$-greedy
- Practice:
  - Sample size $d$ often set to 1

AGARWAL, D., B. LONG, J. TRAUPMAN, D. XIN, AND L. ZHANG (2014): "LASER: A Scalable Response Prediction Platform for Online Advertising," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, p. 173–182, New York, NY, USA. Association for Computing Machinery.

CHAPELLE, O., AND L. LI (2011): "An Empirical Evaluation of Thompson Sampling," in *Advances in Neural Information Processing Systems*, ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, vol. 24. Curran Associates, Inc.

GRAEPEL, T., J. Q. CANDELA, T. BORCHERT, AND R. HERBRICH (2010): "Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine," in *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML '10)*, pp. 13–20. Omnipress, Madison, WI, USA,.

# References II


Hill, D. N., H. Nassif, Y. Liu, A. Iyer, and S. Vishwanathan (2017): "An Efficient Bandit Algorithm for Realtime Multivariate Optimization," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17. ACM.

Russo, D. J., B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. (2018): "A tutorial on thompson sampling," *Foundations and Trends® in Machine Learning*, 11(1), 1–96.

Scott, S. L. (2010): "A modern Bayesian look at the multi-armed bandit," *Applied Stochastic Models in Business and Industry*, 26(6), 639–658.

Scott, S. L. (2015): "Multi-armed bandit experiments in the online service economy," *Applied Stochastic Models in Business and Industry*, 31(1), 37–45.

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

Thompson, W. R. (1935): "On the Theory of Apportionment," *American Journal of Mathematics*, 57(2), 450–456.

# Takeaways

## What is Thompson Sampling?

- Models uncertainty about expected rewards using probability distributions
- Samples from posterior of each arm's reward distribution
- Selects the arm with the highest sampled value
- Posterior is updated after each observation
- Achieves log regret
- Applied at, e.g., Google, Amazon, Facebook, Salesforce, and Netflix

## What is Thompson Sampling?

- Models uncertainty about expected rewards using probability distributions
- Samples from posterior of each arm's reward distribution
- Selects the arm with the highest sampled value
- Posterior is updated after each observation
- Achieves log regret
- Applied at, e.g., Google, Amazon, Facebook, Salesforce, and Netflix (e.g., **??????**)