

Thompson Sampling:

a provably good heuristic for multi-armed
bandits

Shipra Agrawal
Microsoft Research

Contributors: Navin Goyal, Purushottam Kar

Stochastic multi-armed bandits

- Classic model to capture Explore-Exploit conflict
- Applications to clinical trials, online advertising, web search, multi-agent systems, queuing and scheduling
- N arms, unknown reward distributions
 - Expected reward often assumed to be parameterized by some unknown parameter μ
- Pull arms to maximize total reward / minimize regret
 - Minimize regret in expectation or high probability

Thompson Sampling

[W. R. Thompson. *Biometrika*, 1933] Also known as posterior sampling

A simple natural Bayesian heuristic

- Maintain a belief (distribution) for the unknown parameters
- Every time you play an arm and observe a reward, update the belief in Bayesian manner
- At time t , play an arm with its probability of being the best arm, according to the current belief distribution
 - Sample the parameters from the belief distribution.
 - Play the best arm according to these sample parameters

History

- The general principle proposed in [Thompson 1933]
- Rediscovered numerous times independently in the context of reinforcement learning [Wyatt 1997; Ortega & Braun 2010; Strens 2000]
- UCB based algorithms provide good theoretical bounds for many versions of MAB [Auer et al. 2002, ...]
- Thompson Sampling has revived interest
 - Promising empirical performance (e.g., robust to delayed feedback)
[Kaufmann et al. ALT 2012][Chapelle, Li, NIPS 2011],...
 - Used in industrial application [Graepel et al. ICML 2010]
 - Easy to implement: Posterior updates are simple and efficient
 - More on this later

New theoretical guarantees

Thompson Sampling matches the best available guarantees for

- Classic MAB
 - 1-of-N MAB [Agrawal, Goyal COLT 2012], [Kaufmann et al. ALT 2012], [Agrawal, Goyal AISTATS 2013]
 - K-of-N MAB with sub modular rewards [unpublished]
 - MAB with side observations [unpublished]
 - MAB with Delayed feedback [unpublished]
- Contextual MAB
 - **Linear Contextual MAB** [Agrawal Goyal ICML 2013] [New Improvements to match UCB]
 - Sparse Linear Contextual MAB [unpublished joint work with Purushottam Kar]
 - Kernelized Contextual MAB [unpublished joint work with Purushottam Kar]

Linear contextual bandits

- Parameters of the problem:
 - A_t : set of d – dimensional contexts at time t
 - Can be chosen in an adaptive adversarial manner
 - Each context/feature vector in A_t corresponds to an arm (possibly infinite N)
 - $\mu \in R^d$ (**unknown**), $\|\mu\| \leq 1$
- If arm corresponding to context $b(t) \in A_t$ is played at time t then there is a reward $r(t)$:
 - $E[r(t)] = \mu^T b(t), \quad |r(t) - E[r(t)]| \leq R$
- Arm with maximum expected reward: $b^*(t) = \operatorname{argmax}_{b \in A_t} \mu^T b$
- Regret for playing arm $b(t)$ is $\mu^T b^*(t) - \mu^T b(t)$
- Total regret in time T , $R(T) = \sum_t (b_{a^*(t)}(t)^T \mu - b_{a(t)}(t)^T \mu)$

Thompson Sampling

using Gaussian belief distribution

- Start with $N(0_d, v^2 I_d)$ prior belief for unknown parameter μ
- For Gaussian likelihood of rewards, $\Pr(r_t | b(t)^T \mu) \sim N(b(t)^T \mu, v^2)$
- At time t , Bayesian posterior is $N(\hat{\mu}(t+1), v^2 B(t+1)^{-1})$
- Least square estimate
 - $\hat{\mu}(t), B(t)$ given by regularized least square estimate on $(b(1), r(1), \dots, b(t-1), r(t-1))$
 - $B(t) = I_d + \sum_{\tau=1}^{t-1} b(\tau)b(\tau)^T$, $\hat{\mu}(t) = B(t)^{-1} \sum_{\tau=1}^{t-1} b(\tau)r(\tau)$



Gaussian assumption made only for Bayesian interpretation of the algorithm, **regret bounds will hold irrespective of actual reward distribution.**

Thompson Sampling

ALGO

At time t ,

- Sample $\tilde{\mu}(t)$ from $N(\hat{\mu}(t), g_t^2 B(t)^{-1})$
- Play arm $\arg \max_{b \in A_t} b^T \tilde{\mu}(t)$


INITIALIZATION

- $\hat{\mu}(1) = 0, B(1) = I_d, g_t = \sqrt{9Rd \ln(\frac{t}{\delta})}$

Thompson Sampling vs. UCB based algorithms


THOMPSON SAMPLING

At time t ,

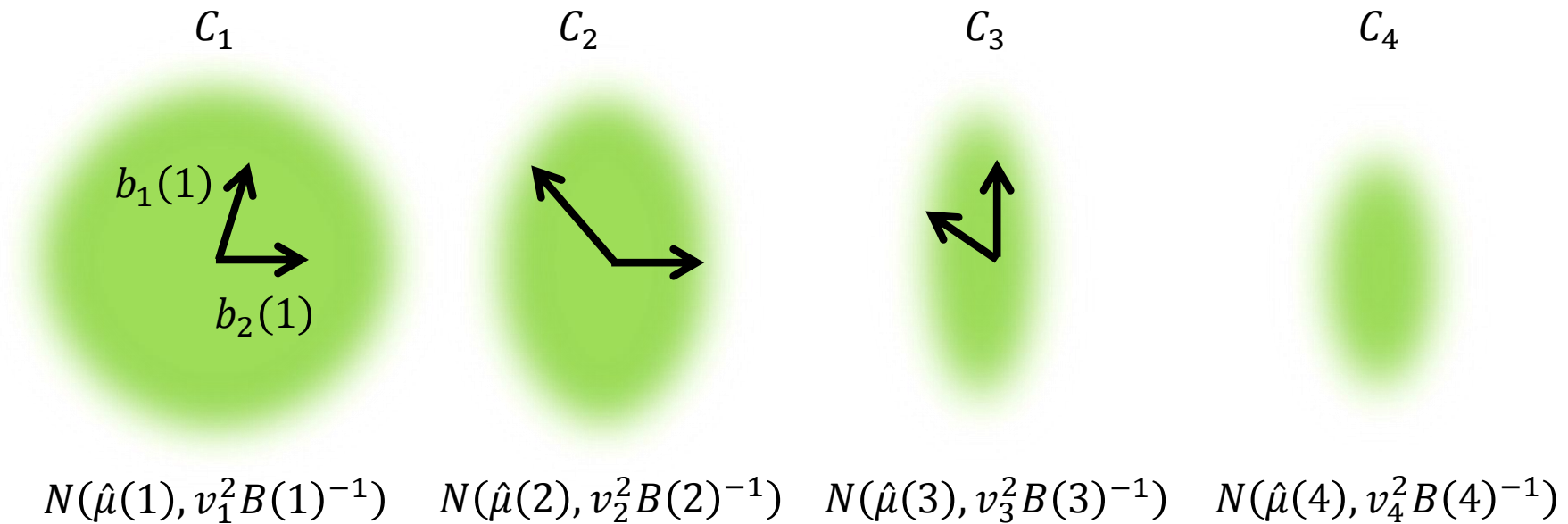
- Sample $\tilde{\mu}(t)$ from $N(\hat{\mu}(t), g_t^2 B(t)^{-1})$
- Play arm $\arg \max_{b \in A_t} b^T \tilde{\mu}(t)$  Randomized version of UCB

UCB based algorithms [Dani et al 2008, Abbasi-Yadkori et al 2011]

At time t ,

- Consider ellipsoid $C_t = (\hat{\mu}(t), g_t^2 B(t)^{-1})$
- Play arm $\arg \max_{b \in A_t} \max_{\tilde{\mu} \in C_t} b^T \tilde{\mu}$  Difficult optimization problem even when A_t is convex

Thompson sampling for linear contextual bandits



Regret bounds

- With probability $1 - \delta$, regret

$$R(T) \leq O\left(d\sqrt{T} \ln(T) \sqrt{\ln\frac{T}{\delta}}\right)$$

- Improvement by a factor d from [Agrawal and Goyal 2013]
- Lower bound is $\Omega(d\sqrt{T})$

- For UCB based algorithms best bound is $O\left(d\sqrt{T} \sqrt{\ln(T) \ln\frac{T}{\delta}}\right)$ [Abbasi-Yadkori et al 2011]

Proof outline

- Regret at time $t = \Delta_{b(t)} = b^*(t)^T \mu - b(t)^T \mu$
- Key Inequality: With high probability (for most histories F_{t-1})

$$E[\Delta_{b(t)} | F_{t-1}] \leq 2\sqrt{\ln T} g_t E[s_{b(t)}(t) | F_{t-1}]$$

- Filtration F_{t-1} : includes history until $t - 1$ and context set A_t
- $s_b(t) = \sqrt{b^T B(t)^{-1} b}$: Can be interpreted as standard deviation in the direction of context b
- $g_t = \sqrt{9R d \ln(t/\delta)}$
- $\sum_{t=1}^T s_{b(t)}(t) = O(\sqrt{dT \ln(T)})$ [Auer 2002, Chu et al 2011]
- Using Azuma-Hoeffding bounds for super-martingales: with probability $1 - \delta$
 - $\sum_t \Delta_{b(t)} \leq O\left(g_t \sqrt{dT \ln(T)}\right) = \tilde{O}(d\sqrt{T})$

Proof techniques

$$C_t = (\hat{\mu}(t), g_t^2 B(t)^{-1})$$

- Lemma: With high prob. the following events happen (for most F_{t-1}), $\mu \in C_t$

$$\forall b \in A_t, |\hat{\mu}(t)^T b - \mu^T b| \leq g_t \sqrt{b^T B(t)^{-1} b} = g_t s_b(t)$$


- From upper confidence bounds for least square estimate in UCB based algorithms [Rusmevichientong-Tsitkilis 2010, Abbasi-Yadkori et al 2011]

- Lemma: For all $F_{t-1}, \forall b \in A_t$,

$$|\tilde{\mu}(t)^T b - \hat{\mu}(t)^T b| \leq \sqrt{\ln(T)} g_t s_b(t)$$

- Variance of Gaussian $\tilde{\mu}$ along context $b = g_t^2 s_b(t)^2 = g_t^2 (b^T B(t)^{-1} b)$

- From now on, assume for all $F_{t-1}, \forall b \in A_t$

 $|\tilde{\mu}(t)^T b - \mu^T b| \leq 2\sqrt{\ln(T)} g_t s_b(t)$

- μ
- $\hat{\mu}(t)$

Proof of key inequality

$$E[\Delta_{b(t)} | F_{t-1}] \leq O(\sqrt{\ln T} g_t) E[s_{b(t)}(t) | F_{t-1}]$$

- Regret at time t , $\Delta_{b(t)} = b^*(t)^T \mu - b(t)^T \mu$
$$\leq (b^*(t)^T \tilde{\mu}(t) - b(t)^T \tilde{\mu}(t)) + \sqrt{\ln T} g_t s_{b^*(t)}(t) + \sqrt{\ln T} g_t s_{b(t)}(t)$$
$$\leq \sqrt{\ln T} g_t s_{b^*(t)}(t) + \sqrt{\ln T} g_t s_{b(t)}(t)$$

Does not go down fast enough unless optimal arm is played often!

ASIDE

- Easy to prove this for UCB.

$$\begin{aligned} \Delta_{b(t)} &= b^*(t)^T \mu - b(t)^T \mu \\ &\leq (b^*(t)^T \mu - b(t)^T \tilde{\mu}(t)) + g_t s_{b(t)}(t) \\ &\leq g_t s_{b(t)}(t) \end{aligned}$$

Proof of key inequality

$$E[\Delta_{b(t)} | F_{t-1}] \leq O(\sqrt{\ln T} g_t) E[s_{b(t)}(t) | F_{t-1}]$$

- Class of Unsaturated (high-variance, low regret) arms at time t

$$b \in A_t: \Delta_b \leq (\sqrt{\ln T} g_t) s_b(t)$$

- Lemma: $\Pr(\text{playing unsaturated arm at time } t | F_{t-1}) \geq p$ (constant)

- Variance is high, mean reward is high -- anti-concentration of posterior ensures that $b^T \tilde{\mu}(t)$ is sufficiently large

- $E[s_{b(t)}(t) | F_{t-1}] \geq p s_{\bar{b}(t)}(t)$

- $\bar{b}(t)$ is the arm with smallest $s_b(t)$ among high variance arms

- What remains to show is that $\Delta_{b(t)} \leq (\sqrt{\ln T} g_t)(s_{\bar{b}(t)}(t) + s_{b(t)}(t))$

Summary: proof of regret bound

$$\begin{aligned}\Delta_{b(t)} &= b^*(t)^T \mu - b(t)^T \mu \\ &\leq \Delta_{\bar{b}(t)} + \bar{b}(t)^T \mu - b(t)^T \mu \\ &\leq \sqrt{\ln T} g_t s_{\bar{b}(t)}(t) + \sqrt{\ln T} g_t s_{\bar{b}(t)}(t) + \sqrt{\ln T} g_t s_{b(t)}(t)\end{aligned}$$

Concentration
Definition of high-variance arms

$$E[\Delta_{b(t)} | F_{t-1}] \leq \left(3\sqrt{\ln T} g_t \frac{1}{c}\right) E[s_{b(t)}(t) | F_{t-1}]$$

$$\sum_t \Delta_{b(t)} \leq O(\sqrt{\ln T} g_t) \sum_t s_{b(t)}(t) \leq O(\sqrt{\ln T} g_t \cdot \sqrt{dT \ln(T)})$$

Conclusion

- Strong modular techniques for analysis of TS
- Thompson Sampling is an attractive option to consider when designing algorithms for new versions of this problem
 - Attractive both theoretically and empirically

Thank you.

Online Learning to Thompson Sampling

[Inspired by Abbasi-Yadkori et al. 2012]

Given: online linear programming algorithm that

- At time t , Inputs $(X_1, Y_1, \dots, X_{t-1}, Y_{t-1}, X_t)$, predicts \hat{Y}_t
- Regret guarantees: with probability $1 - \delta$

$$\sum_t^T \ell(\hat{Y}_t) - \ell(Y_t) \leq h_t$$

- $\ell(y) = (y - \mu^T X_t)^2$

Online Learning to Thompson Sampling

- Start with $N(0_d, I_d)$ belief distribution
- At time t , belief is $N(\hat{\mu}(t), g_t^2 B^{-1}(t))$
 - $\hat{\mu}(t), B(t)$ given by least square estimate on $(b(1), \overset{\hat{Y}_1}{r(1)}, \dots, b(t-1), \overset{\hat{Y}_{t-1}}{r(t-1)})$
 - \hat{Y}_{t-1} is the predictor given by online algo for input $(b(1), r(1), \dots, b(t-1), r(t-1), b(t))$
 - $g_t = h_t$

ALGO: At time t ,

- Sample $\tilde{\mu}(t)$ from $N(\hat{\mu}(t), g_t^2 B^{-1}(t))$
- Play arm $\arg \max_{b \in A_t} b^T \tilde{\mu}(t)$

Regret bounds

- With probability $1 - \delta$, regret

$$R(T) \leq O(h_t \sqrt{dT} \ln(T))$$

- \sqrt{dT} represents the bandit cost over and above the online learning regret
- $h_t = O(\sqrt{d \ln(\frac{T}{\delta})})$ for regularized least square estimate/Follow the regularized leader, to get $\tilde{O}(d\sqrt{T})$ regret
- $h_t = O(\sqrt{s \ln(\frac{T}{\delta})})$ for sparse online learning with sparsity s , to get $\tilde{O}(\sqrt{sdT})$ regret

[For UCB based algorithm OFUL this regret bound was proven in Abbasi-Yadkori et al. 2011]