Thompson Sampling Algorithms for Mean-Variance Bandits

Qiuyu Zhu   Vincent Y. F. Tan

Institute of Operations Research and Analytics,
National University of Singapore

ICML 2020

Stochastic multi-armed bandit

Problem formulation

A **stochastic multi-armed bandit** is a collection of distributions $\nu = (P_1, P_2, \ldots, P_K)$, where K is the number of the arms.
In each period $t \in [T]$:

1. Player picks arm $i(t) \in \mathcal{A}$.
2. Player observes reward $X_{i(t),t} \sim P_{i(t)}$ for the chosen arm.

Learning policy

A policy $\pi : (t, A_1, X_1, \ldots, A_{t-1}, X_{t-1}) \to [K]$ is characterised by,

$$i(t) = \pi(t, i(1), X_{i(1),1}, \cdots, i(t-1), X_{i(t-1),t-1}), \quad t = 1, \cdots, T$$

The player can only use the past observations in current decisions.

The learning objective

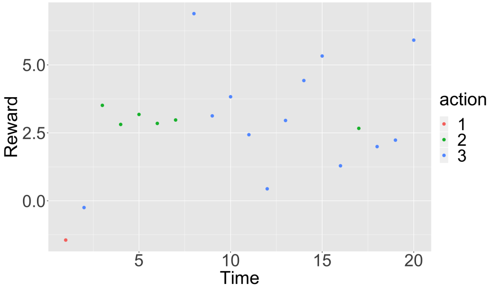
Objective

Minimize the expected cumulative regret

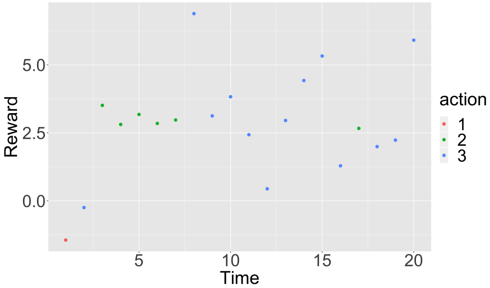
$$\mathcal{R}_n = \mathbb{E}\left[\sum_{t=1}^{n}(X_{i^*,t} - X_{i(t),t})\right] = \sum_{t=1}^{n}(\mu^* - \mu_{i(t)}) = \sum_{i=1}^{K}\Delta_i \mathbb{E}[T_{i,n}]$$

where μ_i is the mean of each arm, $i^* = \arg\max[\mu_i]$, $\Delta_i = \mu^* - \mu_i$ and $T_{i,n} = \sum_{t=1}^{n} \mathbb{1}_{\{i(t)=i\}}$

Motivation

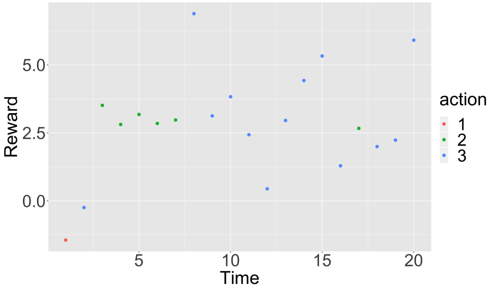


Motivation



- $\text{Mean} = (-1.44, 3.00, 3.12)$

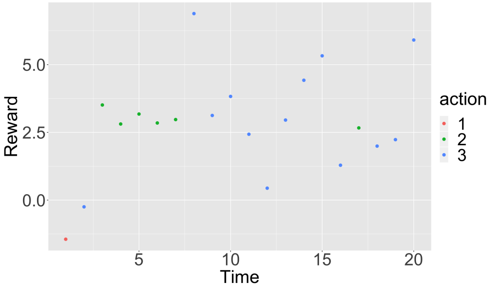
Motivation



True reward distribution:

- Arm 1 $\sim \mathcal{N}(1, 3)$
- Arm 2 $\sim \mathcal{N}(3, 0.1)$
- Arm 3 $\sim \mathcal{N}(3.3, 4)$

Motivation



True reward distribution:

- Arm 1 $\sim \mathcal{N}(1, 3)$
- Arm 2 $\sim \mathcal{N}(3, 0.1)$
- Arm 3 $\sim \mathcal{N}(3.3, 4)$

Some applications require a trade-off between risk and return.

Mean-variance multi-armed bandit

Definition 1 (Mean-Variance)

The mean-variance of an arm i with mean μ_i, variance σ_i^2 and coefficient absolute risk tolerance $\rho > 0$ is defined as

$$\mathrm{MV}_i = \rho\mu_i - \sigma_i^2$$

Definition 2 (Empirical Mean-Variance)

Suppose we have i.i.d. samples $\{X_{i,t}\}_{t=1}^{s}$ from the distribution ν_i, the empirical mean-variance is defined as

$$\widehat{\mathrm{MV}}_{i,s} = \rho\hat{\mu}_{i,s} - \hat{\sigma}_{i,s}^2$$

where $\hat{\sigma}_{i,s}^2$ and $\hat{\mu}_{i,s}$ are empirical variance and mean respectively.

The learning objective

For a given policy π, and its corresponding performance over n rounds $\{Z_t, t = 1, 2, \ldots, n\}$. We define its empirical mean-variance as

$$\widehat{\mathrm{MV}}_n(\pi) = \rho\hat{\mu}_n(\pi) - \hat{\sigma}_n^2(\pi)$$

where

$$\hat{\mu}_n(\pi) = \frac{1}{n}\sum_{t=1}^{T} Z_t, \quad \text{and} \quad \hat{\sigma}_n^2(\pi) = \frac{1}{n}\sum_{t=1}^{n}(Z_t - \hat{\mu}_n(\pi))^2.$$

Definition 3 (Regret)

The expected regret of a policy $\pi(\cdot)$ over n rounds is defined as

$$\mathbb{E}[\mathcal{R}_n(\pi)] = n\left(\mathrm{MV}_1 - \mathbb{E}\left[\widehat{\mathrm{MV}}_n(\pi)\right]\right)$$

where we assume the first arm is the best arm.

The variances

Law of total variance

$$Var(reward) = \mathbb{E}[Var(reward|arm)] + Var(\mathbb{E}[reward|arm])$$

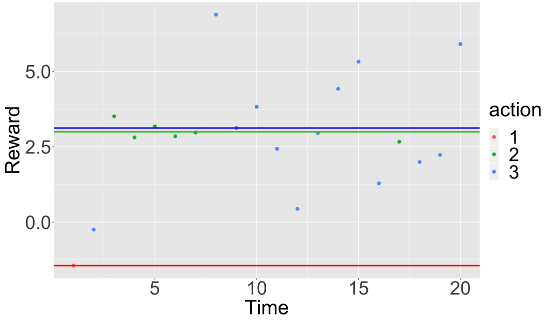


Figure 1: Reward Process

Pseudo-regret

Definition 4

The expected pseudo-regret for a policy $\pi(\cdot)$ over n rounds is defined as

$$\mathbb{E}\big[\widetilde{\mathcal{R}}_n(\pi)\big] = \sum_{i=2}^{K} \mathbb{E}\left[T_{i,n}\right] \Delta_i + \frac{1}{n} \sum_{i=1}^{K} \sum_{j \neq i} \mathbb{E}\left[T_{i,n} T_{j,n}\right] \Gamma_{i,j}^2.$$

where $\Delta_i = \sigma_i^2 - \sigma_1^2 - \rho(\mu_i - \mu_1)$ is the gap between MV_i and MV_1, and $\Gamma_{i,j}$ is the gap between μ_i and μ_j.

Lemma 1

The difference between the expected regret and the expected pseudo-regret can be bounded as follows:

$$\mathbb{E}\left[\mathcal{R}_n(\pi)\right] \leq \mathbb{E}\big[\widetilde{\mathcal{R}}_n(\pi)\big] + 3 \sum_{i=1}^{K} \sigma_i^2$$

Pseudo-regret

Simplification of pseudo-regret

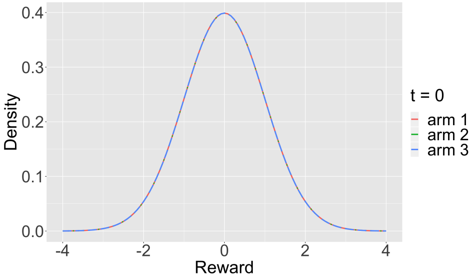
$$\frac{1}{n}\sum_{i=1}^{K}\sum_{j\neq i}\mathbb{E}\left[T_{i,n}T_{j,n}\right]\Gamma_{i,j}^{2}\leq 2\sum_{i=2}^{K}\mathbb{E}\left[T_{i,n}\right]\Gamma_{i,\max}^{2} \tag{1}$$

where $\Gamma_{i,\max}^2 = \max\{(\mu_i - \mu_j)^2 : j = 1, \ldots, K\}$.

By applying Definition 4, Lemma 1 and Eqn. (1), it suffices to bound the expected number of pulls of suboptimal arms $\mathbb{E}[T_{i,n}]$.

Thompson Sampling

True reward distributions are: $\mathcal{N}(1,3), \mathcal{N}(3,0.1), \mathcal{N}(3.3,4)$



$t = 0$

$\rightarrow$ Samples: $(1.30, 1.22, -0.07)$

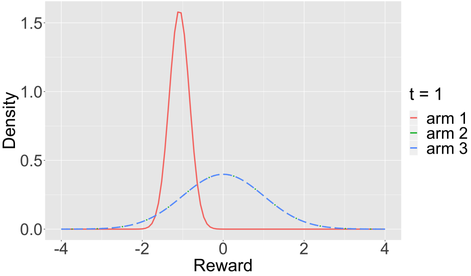
$\rightarrow$ Play arm 1

$\rightarrow$ Get reward -1.44

Update posteriors

Thompson Sampling

True reward distributions are: $\mathcal{N}(1,3), \mathcal{N}(3,0.1), \mathcal{N}(3.3,4)$



$t = 1$

$\rightarrow$ Samples: $(0.17, -0.24, 0.65)$

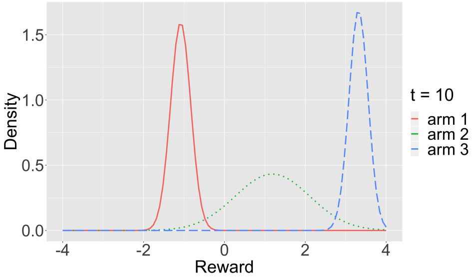
$\rightarrow$ Play arm 3

$\rightarrow$ Get reward 0.62

Update posteriors

Thompson Sampling

True reward distributions are: $\mathcal{N}(1, 3), \mathcal{N}(3, 0.1), \mathcal{N}(3.3, 4)$



$t = 10$

$\rightarrow$ Samples: $(-0.24, 2.15, 3.23)$

$\rightarrow$ Play arm 2

$\rightarrow$ Get reward 2.12

Update posteriors

TS algorithm for mean learning

Algorithm 1 Thompson Sampling for Mean Learning

Input: $\hat{\mu}_{i,0} = 0$, $T_{i,0} = 0, \alpha_{i,0} = \frac{1}{2}, \beta_{i,0} = \frac{1}{2}$.
for each $t = 1, 2 \ldots,$ **do**
  Sample $\theta_i(t)$ from $\mathcal{N}(\hat{\mu}_{i,t-1}, 1/(T_{i,t-1} + 1))$.
  Play arm $i(t) = \arg\max_i \rho\theta_i(t) - 2\beta_{i,t-1}$ and observe $X_{i(t),t}$
  $(\hat{\mu}_{i(t),t}, T_{i(t),t}, \alpha_{i(t),t}, \beta_{i(t),t}) =$
    $\text{Update}(\hat{\mu}_{i(t),t-1}, T_{i(t),t-1}, \alpha_{i(t),t-1}, \beta_{i(t),t-1}, X_{i(t),t})$
end for

Regret bound

Theorem 1

If $\rho > \max\left\{\sigma_1^2/\Gamma_i : i = 1, 2, \ldots, K\right\}$, *the asymptotic expected regret incurredd by MTS for mean-variance Gaussian bandits satisfies*

$$\overline{\lim_{n\to\infty}} \frac{\mathbb{E}\left[\widetilde{\mathcal{R}}_n(\mathrm{MTS})\right]}{\log n} \leq \sum_{i=2}^{K} \frac{2\rho^2}{\left(\rho\Gamma_{1,i} - \sigma_1^2\right)^2} \left(\Delta_i + 2\Gamma_{i,\max}^2\right)$$

Remark 1 (The bound)

Since $\Delta_i = \sigma_i^2 - \sigma_1^2 + \rho\Gamma_{1,i}$, *as* ρ *tends to* $+\infty$, *we observe that*

$$\overline{\lim_{n\to\infty}} \frac{\mathbb{E}\left[\widetilde{\mathcal{R}}_n(\mathrm{MTS})\right]}{\rho \log n} \leq \sum_{i=2}^{K} \frac{2}{\Gamma_{1,i}}.$$

This bound is near-optimal according to [Agrawal and Goyal, 2012].

TS algorithm for variance learning

Algorithm 2 TS for Variance Learning

Input: $\hat{\mu}_{i,0} = 0, T_{i,0} = 0, \alpha_{i,0} = \frac{1}{2}, \beta_{i,0} = \frac{1}{2}$.
for each $t = 1, 2 \ldots,$ **do**
  Sample $\tau_i(t)$ from $\mathrm{Gamma}\left(\alpha_{i,t-1}, \beta_{i,t-1}\right)$.
  Play arm $i(t) = \arg\max_{i \in [K]} \rho \hat{\mu}_{i,t-1} - 1/\tau_i\left(t\right)$ and observe $X_{i(t),t}$
  $(\hat{\mu}_{i(t),t}, T_{i(t),t}, \alpha_{i(t),t}, \beta_{i(t),t}) =$
  $\mathrm{Update}(\hat{\mu}_{i(t),t-1}, T_{i(t),t-1}, \alpha_{i(t),t-1}, \beta_{i(t),t-1}, X_{i(t),t})$
end for

Regret bound

Theorem 2

Let $h(x) = \frac{1}{2}(x - 1 - \log x)$*. If* $\rho \le \min\left\{\Delta_i/\Gamma_i : \Delta_i/\Gamma_i > 0\right\}$*, the asymptotic regret incurred by VTS for mean-variance Gaussian bandits satisfies*

$$\overline{\lim_{n\to\infty}} \frac{\mathbb{E}\left[\widetilde{\mathcal{R}}_n(\mathrm{VTS})\right]}{\log n} \le \sum_{i=2}^{K} \frac{1}{h\left(\sigma_i^2/\sigma_1^2\right)} \left(\Delta_i + 2\Gamma_{i,\max}^2\right).$$

Remark 2 (Order optimality)

Vakili and Zhao (2015) proved that the expected regret of any consistent algorithm is $\Omega\left((\log n)/\Delta^2\right)$ *where* $\Delta = \min_{i\neq 1} \Delta_i$*. Since*

$$h(x) = (x-1)^2/4 + o((x-1)^2) \text{ as } x \to 1,$$

MTS and VTS are order optimal in both n *and* Δ*.*

TS algorithm for mean-variance learning

Algorithm 3 Thompson Sampling for Mean-Variance bandits (MVTS)

Input: $\hat{\mu}_{i,0} = 0$, $T_{i,0} = 0, \alpha_{i,0} = \frac{1}{2}, \beta_{i,0} = \frac{1}{2}$.
for each $t = 1, 2, \ldots,$ **do**
  Sample $\tau_i(t)$ from $\mathrm{Gamma}(\alpha_{i,t-1}, \beta_{i,t-1})$.
  Sample $\theta_i(t)$ from $\mathcal{N}(\hat{\mu}_{i,t-1}, 1/(T_{i,t-1} + 1))$
  Play arm $i(t) = \arg\max_{i \in [K]} \rho\theta_i(t) - 1/\tau_i(t)$ and observe $X_{i(t),t}$
  $(\hat{\mu}_{i(t),t}, T_{i(t),t}, \alpha_{i(t),t}, \beta_{i(t),t}) =$
    $\mathrm{Update}(\hat{\mu}_{i(t),t-1}, T_{i(t),t-1}, \alpha_{i(t),t-1}, \beta_{i(t),t-1}, X_{i(t),t})$
end for

Hierarchical structure of Thompson samples

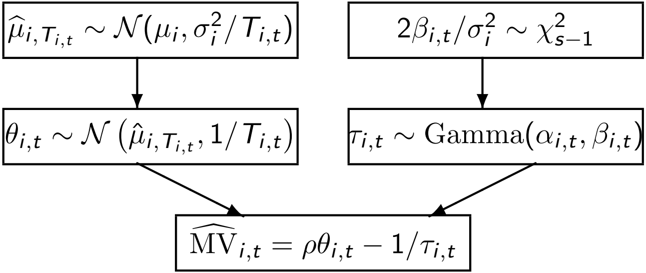


Figure 2: Hierarchical structure of the mean-variance Thompson samples in MVTS.

Regret bound

Theorem 3

The asymptotic expected regret of MVTS for mean-variance Gaussian bandits satisfies

$$\overline{\lim_{n\to\infty}} \frac{\mathbb{E}\left[\widetilde{\mathcal{R}}_n(\mathrm{MVTS})\right]}{\log n} \leq \sum_{i=2}^{K} \max\left\{\frac{2}{\Gamma_{1,i}^2}, \frac{1}{h(\sigma_i^2/\sigma_1^2)}\right\}\left(\Delta_i + 2\Gamma_{i,\max}^2\right).$$

Remark 3

Regret bound of MVTS particularizes to MTS and VTS when $\rho \to \infty$ and $\rho \to 0^+$ respectively.

Hence, MVTS is order optimal when ρ assumes these extremal values.

Numerical Simulations

MV-LCB is the algorithm from [Sani et al., 2012],[Vakili and Zhao, 2016].

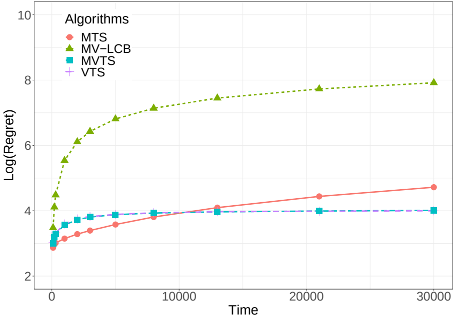


Figure 3: $\rho = 10^{-3}$

The $K = 15$ Gaussian arms are set to the same as the experiments from Sani et al. [2012] (i.e. $\mu = (0.1, 0.2, \ldots, 0.79)$, $\sigma_i^2 = (0.05, 0.34, \ldots, 0.85)$)

Numerical Simulations

MV-LCB is the algorithm from [Sani et al., 2012],[Vakili and Zhao, 2016].

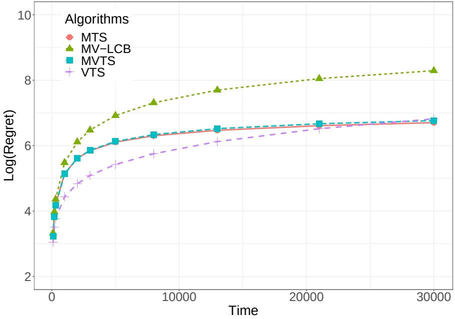


Figure 4: $\rho = 1$

The $K = 15$ Gaussian arms are set to the same as the experiments from Sani et al. [2012] (i.e. $\mu = (0.1, 0.2, \ldots, 0.79)$, $\sigma_i^2 = (0.05, 0.34, \ldots, 0.85)$)

Numerical Simulations

MV-LCB is the algorithm from [Sani et al., 2012],[Vakili and Zhao, 2016].

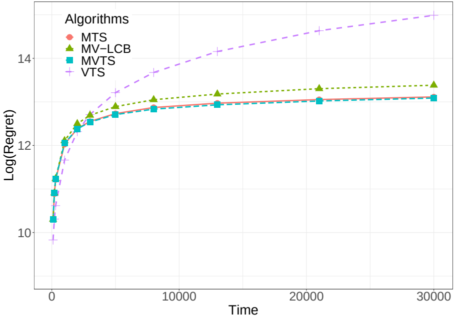


Figure 5: $\rho = 1000$

The $K = 15$ Gaussian arms are set to the same as the experiments from Sani et al. [2012] (i.e. $\mu = (0.1, 0.2, \ldots, 0.79)$, $\sigma_i^2 = (0.05, 0.34, \ldots, 0.85)$)

Numerical Simulations

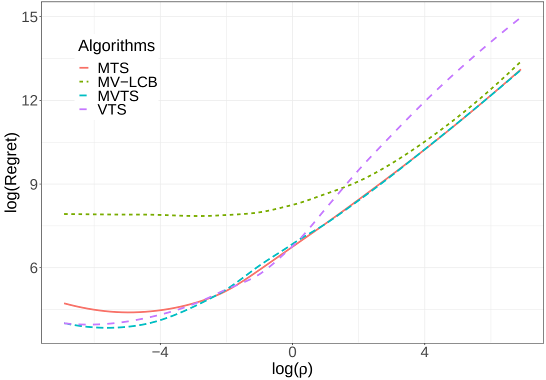


Figure 6: Regret of Gaussian MV MAB with $K = 15$

Thank you for listening!

Q&A