# Thompson Sampling

Source: Bandit Algorithms, T. Lattimore and C. Szepesvari, Cambridge Univ. Press, 2019 (to appear); Chapter 36.

Thompson sampling is a Bayesian bandit algorithm for a k-armed bandit.

Setting: $(\mathcal{E}, \mathcal{G}, Q, P)$ → reward

- $\mathcal{E}$: parameter space; typically a k-length mean vector
- $Q$: prior on $\nu \in \mathcal{E}$

Example: k-armed bandit, with Gaussian rewards and Gaussian prior. $\mathcal{E} = \mathbb{R}^k$, $\mathcal{G} = \sigma(\mathbb{R}^k)$ = Borel-$\sigma$-algebra over $\mathbb{R}^k$, $Q$ a prior over $\nu \in \mathcal{E}$: $\nu = (\nu_1, \nu_2 \cdots \nu_k)^T$.

The interpretation is: $\nu_k$ is the (unknown) mean reward for arm k. This has a generative model: $\nu \sim \mathcal{E}_N^k(1)$, i.e.,

$$Q = N(0, I); \quad P = (P_{k\nu},\ k = 1, 2, \cdots, k,\ \nu \in \mathcal{E}),$$

↳ prior is '0' mean, unit variance for all arms

i.e., $P_{k\nu} = N(\nu_k, 1)$. An arm pull of the $k^{th}$ arm generates a reward r.v. $X$, where $X$ has a conditional dist. (given $\nu$) of $X \sim P_{k\nu}$.

## Thompson Sampling Algorithm:

(1) Given samples $X_1, X_2, \ldots, X_{t-1}$, update the prior dist. on the environment, i.e., maintain posterior dist: $Q(\cdot \mid X_1, \ldots X_{t-1})$, using Baye's rule starting from the prior dist. $Q(\cdot)$.

(2) Draw a sample from this posterior dist: Sample an instance $\nu_t = (\nu_{1,t}, \nu_{2,t} \ldots \nu_{k,t})^\top \sim Q(\cdot \mid X_1, \ldots, X_{t-1})$

(3) Play arm using the sample $\nu_t = (\nu_{1,t} \ldots \nu_{k,t})^\top$:

$$A_t = \underset{1 \le j \le k}{\operatorname{argmax}} \ \nu_{j,t}$$

and break ties as the genie policy (which knows the environment instance).

---

Note 1: The textbook uses a more general notation. Specifically, the environment $\nu \in \mathcal{E}$ could include additional parameters, and the mean of the $j^{th}$ arm under environment $\nu$ is denoted by $\mu_j(\nu)$. In these notes, we are instead simply using $\nu_j$ or $\nu_{j,t}$ if we need to index both

an arm $j$ and a time $t \in \{1, 2, \cdots, n\}$.

---

Note 2: We have a doubly stochastic process here. Specifically, $\nu_t \sim Q(\cdot \mid X_1, \cdots, X_{t-1})$ is a $k$-dim. random vector, and the reward $X_t$ is a r.v.; given that $\{A_t = j\} \cap \{\nu_t = \nu_t\}$, $X_t \mid (j, \nu_t) \sim N(\nu_{j,t}, 1)$.

the r.v. corresp. to the environment at time $t$

---

Note 3: A key property of Thompson sampling is that the conditional prob. of playing arm $j$ under Thompson sampling is the same as that of the optimal (genie) policy. Formally,

genie

$$P\left(A_t^* = j \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}\right) \longrightarrow (1)$$

player

$$= P\left(A_t = j \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}\right)$$

* This does not mean $X_{\{A_t^* = j\}} = X_{\{A_t = j\}}$ !

* Observe that the above does not have any conditioning w.r.t. the environment sample $\nu$ that was chosen

at $t=0$. This environment $\nu$ remains invariant for any $t=1,2,\cdots,n$ along this sample path. The genie knows this, and thus $\{A_t^* = j\}$ is a function of $\nu$ (specifically, $A_t^* = \underset{1 \le i \le k}{\text{argmax}}\ \nu_i$).

Thus to compute $P(A_t^* = j \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1})$, we further need to condition and uncondition on the environment. Also finally note that the above property holds only for the marginal dist. of the arm pull.

Pf of ①: $\mathcal{F}_{t-1} = \sigma\{A_1, X_1, \cdots, A_{t-1}, X_{t-1}\}$.

$$P(A_t^* = j \mid \mathcal{F}_{t-1}) = \int_{\nu \in \mathcal{E}} P(A_t^* = j \mid \mathcal{F}_{t-1}, \nu)\, dP(\nu \mid \mathcal{F}_{t-1})$$

↳ conditioning on the true environment

But $P(\nu \in E \mid \mathcal{F}_{t-1}) = Q(\nu \in E \mid \mathcal{F}_{t-1})$

This follows because the true environment $\nu \sim Q(\cdot)$, and the update given $\mathcal{F}_{t-1}$ is the posterior as above.

Ⓐ

$$\Rightarrow P(A_t^* = j \mid \mathcal{F}_{t-1}) = \int_{\nu \in \mathcal{E}} P(A_t^* = j \mid \mathcal{F}_{t-1}, \nu) \cdot dQ(\nu \mid \mathcal{F}_{t-1})$$

$\delta\{\nu_j = \max_{1 \le i \le k} \nu_i\}$.

(for simplicity, we have assumed unique max.; else use tie-breaking rule)

$$P(A_t = j \mid \mathcal{F}_{t-1}) = \int_{\nu_t \in \mathcal{E}} P(A_t = j \mid \mathcal{F}_{t-1}, \nu_t) \, dP(\nu_t \mid \mathcal{F}_{t-1})$$

↳ conditioning on the sample drawn in Thompson sampling algorithm.

But by defn of Thompson sampling,

$$\nu_t \sim Q(\cdot \mid \mathcal{F}_{t-1})$$

$$\Rightarrow P(\nu_t \mid \mathcal{F}_{t-1}) = Q(\nu_t \mid \mathcal{F}_{t-1}) \quad (B)$$

$$\therefore P(A_t = j \mid \mathcal{F}_{t-1}) = \int_{\nu_t \in \mathcal{E}} P(A_t = j \mid \mathcal{F}_{t-1}, \nu_t) \cdot dQ(\nu_t \mid \mathcal{F}_{t-1})$$

$$\delta\{\nu_{j,t} = \max_{1 \le i \le k} \nu_{i,t}\}$$

Now interpreting $\nu$ and $\nu_t$ as dummy variables in (A), (B), we observe these expressions are the same. ☑

---

## Analyzing Bayesian Regret for Thompson Sampling

$$BR_n(\pi, Q) = E_{\nu \sim Q, P_\nu, R}\left[\sum_{t=1}^{n} (X_{A^*} - X_{A_t})\right]$$

randomness introduced by algorithm

environment ↔ $\nu$

$P_\nu, R$ → arm reward conditioned on $\nu$

$$= E_{\nu \sim Q, R}\left[ n\nu_{A^*} - \sum_{t=1}^{n} \nu_{A_t} \right]$$

$$= E_{\nu \sim Q}\left[ \sum_{t=1}^{n} E_R\left[ \nu_{A^*} - \nu_{A_t} \mid \mathcal{F}_{t-1} \right] \right]$$

**Theorem (36.1 in text):** $(\mathcal{E}, \mathcal{G}, Q, \mathbb{P})$ a Bayesian bandit, with $\{P_{\nu j},\ j = 1, 2, \cdots, k;\ \nu \in \mathcal{E}\}$ uniformly 1-subGaussian, with $\nu_i \in [0,1]$ $\forall\ i, \nu$. Then, Thompson sampling satisfies:

$$BR_n(\pi, Q) \leq C\sqrt{kn \ln(n)}, \qquad C > 0.$$

**Proof:** For $j = 1, 2, \cdots, k$, and $t = 1, 2, \ldots, n$, let

$$U_t(j) = \left[ \hat{\nu}_j(t-1) + \sqrt{\frac{2\ln(1/\delta)}{\max\{1, T_j(t-1)\}}} \right]_0^1$$

↳ clip and project into $[0,1]$.

(Note: $\hat{\nu}_j(\cdot) = 0$ if $T_j(\cdot) = 0$).

$\hat{\nu}_j(t-1)$ = empirical estimate; $\nu_j$ = true value of mean in environment $\nu \in \mathcal{E}$.

$$B = \left\{ \forall\ j \in \{1, \ldots, k\} \text{ and } t \in \{1, 2, \ldots, n\},\ |\hat{\nu}_j(t-1) - \nu_j| < \sqrt{\frac{2\ln(1/\delta)}{\max\{1, T_j(t-1)\}}} \right\}$$

Now, recall $P_{vj} \sim 1\text{-subG}$. Using an union bound over number of samples $T_j(t-1) \in \{1, 2, \ldots, n\}$ and union bound over arms $j = 1, 2, \ldots, k$ (very similar to the corresponding step in the UCB algorithm), we have

$$P(B^c) \leq nk\delta \longrightarrow (2)$$

Next, recall: $\longrightarrow (3)$.

$$BR_n(\pi, Q) = \mathbb{E}_{\nu \sim Q}\left[\sum_{t=1}^{n} \mathbb{E}_R\left[\nu_{A^*} - \nu_{A_t} \mid \mathcal{F}_{t-1}\right]\right]$$

environment ← $\nu \sim Q$; algo. ⇢ $\mathbb{E}_R$

Thus, $\mathcal{F}_{t-1} = \sigma\{A_1, X_1, \ldots, A_{t-1}, X_{t-1}\}$.

$$\mathbb{E}_R\left[\nu_{A^*} - \nu_{A_t} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_R\left[\nu_{A^*} - U_t(A_t) + U_t(A_t) - \nu_{A_t} \mid \mathcal{F}_{t-1}\right]$$

Now $\mathbb{E}_R\left[U_t(A_t) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_R\left[U_t(A^*) \mid \mathcal{F}_{t-1}\right]$.

This follows because $P(A_t = j \mid \mathcal{F}_{t-1}) = P(A^* = j \mid \mathcal{F}_{t-1})$, which was discussed above in Note 3, (1).

$$\therefore \mathbb{E}_R\left[\nu_{A^*} - \nu_{A_t} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_R\left[\nu_{A^*} - U_t(A^*) + U_t(A_t) - \nu_{A_t} \mid \mathcal{F}_{t-1}\right]$$

$\hookrightarrow (4)$.

Substituting ④ in ③, we have:

$$BR_n(\pi, Q) = \mathbb{E}_{\nu \sim Q}\left[\sum_{t=1}^{n} \mathbb{E}_R\left[\nu_{A^*} - U_t(A^*) + U_t(A_t) - \nu_{A_t} \mid \mathcal{F}_{t-1}\right]\right]$$

iterated expectations

$$= \mathbb{E}_{\nu \sim Q, R}\left[\sum_{t=1}^{n}\left(\nu_{A^*} - U_t(A^*)\right) + \sum_{t=1}^{n}\left(U_t(A_t) - \nu_{A_t}\right)\right]$$

$$= \mathbb{E}_{\nu \sim Q, R}\left[\underbrace{\sum_{t=1}^{n}\left(\nu_{A^*} - U_t(A^*) + U_t(A_t) - \nu_{A_t}\right)\chi_{B^c}}_{\text{I}} + \underbrace{\left(\sum_{t=1}^{n}\left(\nu_{A^*} - U_t(A^*)\right)\right)\chi_B}_{\text{II}} + \underbrace{\left(\sum_{t=1}^{n}\left(U_t(A_t) - \nu_{A_t}\right)\right)\chi_B}_{\text{III}}\right]$$

(I) $\leq 2n$ (because all terms $\in [0,1]$).

(II) $\leq 0$

because $B \Rightarrow \nu_{A^*} \leq \hat{\nu}_{A^*} + \sqrt{\frac{2\ln(1/\delta)}{\max\{1, T_j(t-1)\}}}$

Now $U_t(A^*) = \left[\hat{\nu}_{A^*} + \sqrt{\frac{2\ln(1/\delta)}{\max\{1, T_j(\cdot)\}}}\right]_0^1$

If term inside $> 1$, immediately true that (II) $\leq 0$.

If term inside $< 0$, $U_t(A^*) \geq \hat{\nu}_{A^*} + \sqrt{\frac{2\ln(1/\delta)}{\max\{1, T_j(\cdot)\}}}$

else, $U_t(A^*) = \hat{\nu}_{A^*} + \sqrt{\frac{2\ln(1/\delta)}{\max\{1, T_j(\cdot)\}}}$

In either case, $B \Rightarrow \nu_{A^*} - U_t(A^*) \leq 0$.

(III) $= \mathbb{E}_{\nu \wedge Q, R}\left[ \sum_{t=1}^{n} \left( U_t(A_t) - \nu_{A_t} \right) \chi_B \right]$

$= \mathbb{E}_{\nu, R}\left[ \sum_{t=1}^{n} \sum_{j=1}^{k} \chi_{\{A_t = j\}} \left( U_t(k) - \nu_k \right) \chi_B \right]$

$\leq \mathbb{E}_{\nu \wedge Q, R}\left[ \sum_{j=1}^{k} \sum_{t=1}^{n} \chi_{\{A_t = j\}} \sqrt{\frac{2 \ln(1/\delta)}{\max\{1, T_j(t-1)\}}} \chi_B \right]$

$\leq \mathbb{E}_{\nu \wedge Q, R}\left[ \sum_{j=1}^{k} \sum_{t=1}^{n} \chi_{\{A_t = j\}} \sqrt{\frac{2 \ln(1/\delta)}{\max\{1, T_j(t)\}}} \right]$

$\leq \mathbb{E}_{\nu \wedge Q, R}\left[ \sum_{j=1}^{k} \int_{0}^{T_j(n)} \sqrt{\frac{2 \ln(1/\delta)}{s}} \, ds \right]$

$= \mathbb{E}_{\nu \wedge Q, R}\left[ \sum_{j=1}^{k} \sqrt{4 \ln(1/\delta) \, T_j(n)} \right]$

$\sum_j T_j(n) = n$

$\leq \mathbb{E}_{\nu \wedge Q, R}\left[ \sqrt{4 \ln(1/\delta) \, nk} \right]$

deterministic $\Rightarrow$ drop $\mathbb{E}[\cdot]$.

Choose $\delta = 1/n^2$,

and result follows. ∎

# Frequentist (environment dependent) analysis of Thompson Sampling:

**Algorithm:** k-arm setting

Initial: $\{F_1(1), F_2(1), \cdots, F_k(1)\}$ prior CDF for the k arms.

time = 1

$F_1(1)$ → arm 1; $F_k(1)$ → arm = k

For each $t = 1, 2, \cdots, n$:

① Sample possible environment: Choose $\theta_i(t) \sim F_i(t)$, $i = 1, 2, \cdots, k$

② Action: $A_t = \arg\max_i \theta_i(t)$

③ Observe / Update: Observe reward $X_t$ (corresponding to arm $A_t$).

Update $F_{A_t}(t+1) = \text{Update}(F_{A_t}(t), A_t, X_t)$

→ Bayesian posterior update for TS

$F_i(t+1)(x) = P(X_i \le x \mid \mathcal{F}_t)$ $\quad F_i(t+1) = F_i(t)$ $i \neq A_t$

---

**Notation:** $F_{is}(\cdot)$ = Posterior CDF of arm $i$ using

arm → $i$; # of samples → $s$

S samples. (We will use this in an appropriate union bound as in UCB proof to switch from $\{T_i(t-1), t=1,2,\dots,n\}$ to samples $\{s=1,2,\dots,n\}$.)

---

**Theorem (3b.2 in text):** for arm $i$, and $\varepsilon \in \mathbb{R}$,

$$E[T_i(n)] \le 1 + E\left[\sum_{s=0}^{n-1}\left(\frac{1}{G_{is}} - 1\right)\right] + E\left[\sum_{s=0}^{n-1} \chi_{\{G_{is} > 1/n\}}\right]$$

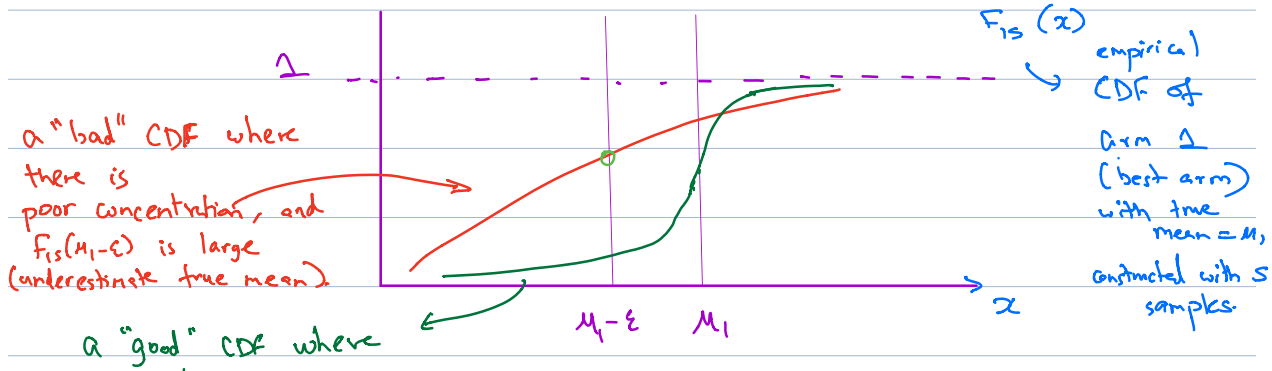
expectation only over randomness of arm reward and algorithm; NOT over environment

where $G_{is} = 1 - F_{is}(\mu_1 - \varepsilon)$.

---

**Intuition:** (assume $\varepsilon > 0$, a small +ve constant)

① $$\left(\frac{1}{G_{is}} - 1\right) = \frac{1 - G_{is}}{G_{is}} = \left(\frac{F_{is}(\mu_1 - \varepsilon)}{1 - F_{is}(\mu_1 - \varepsilon)}\right)$$



The first term $E\left[\sum_{s=0}^{n-1}\left(\frac{1}{G_{is}}-1\right)\right]$ determines the regret due to underestimation of the true mean $\mu_1$. This can cause sub-optimal arms to be played too frequently thus leading to larger regret.

We will later show that in a Gaussian setting, and with appropriate priors s.t. $F_i(t) \sim N\left(\hat{\mu}_i(t), \frac{1}{t}\right)$, and $\mu_1 > \mu_2 \geq \cdots \mu_K$, we have:

$$E\left[\sum_{s=0}^{n-1}\left(\frac{1}{G_{is}}-1\right)\right] \leq \frac{c}{\varepsilon^2}\ln\left(\frac{1}{\varepsilon}\right)$$

(problem 36.6 @ in text)

(2) $$\chi_{\{G_{is} > \frac{1}{n}\}} = \chi_{\{1-F_{is}(\mu_1-\varepsilon) > \frac{1}{n}\}}$$

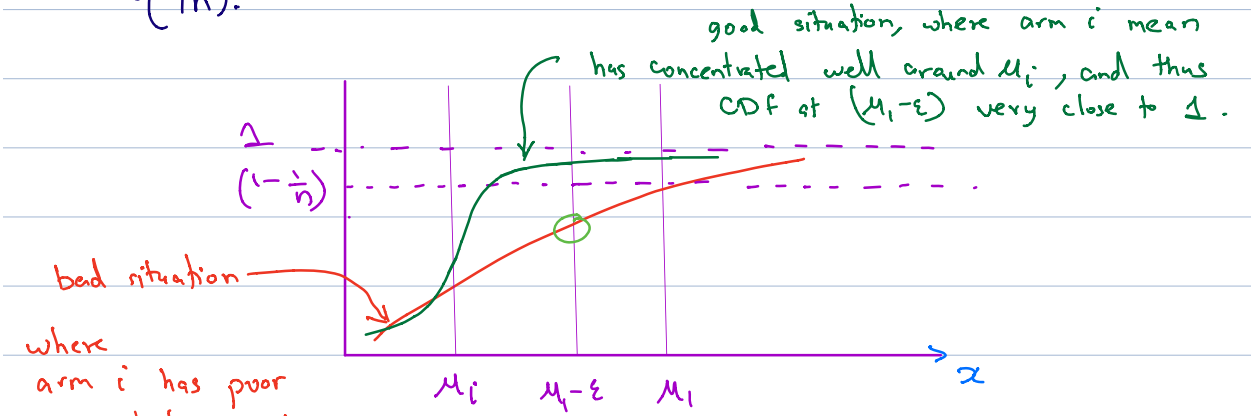
$$= \chi_{\{F_{is}(\mu_1-\varepsilon) < 1-\frac{1}{n}\}}$$

note difference. $i$ vs. $1$.

This controls the overestimation of the $i^{th}$ arm (which is suboptimal), when compared to the mean of the best arm $\mu_1$. Ideally, $F_{is}(\mu_i + \tilde{\varepsilon}) \approx 1$ (i.e., the $i^{th}$ arm's

↳ some small $\tilde{\varepsilon}$, unrelated to $\varepsilon$

mean is estimated correctly). With $\mu_i < \mu_i + \tilde{\varepsilon} < \mu_1 - \varepsilon < \mu_1$, we thus want to be in a regime where we have $F_{is}(\mu_1 - \varepsilon) \approx 1$, where the approx. is extremely small, i.e., $o(1/n)$.




In the same Gaussian setting as ① above, we will later show that: (problem 36.6 ⑤ in text)

$$E\left[\sum_{s=0}^{n-1} \chi\left\{G_{is} > \frac{1}{n}\right\}\right] \leq 1 + \frac{2}{(\Delta_i - \varepsilon)^2}\left(\ln(n) + \sqrt{\pi \ln(n)} + 1\right).$$

---

Note 1: Explicit Expressions in a Gaussian setting.

Suppose that we are in a Gaussian prior - Gaussian rewards

setting, i.e., $Q_i \sim N(\mu_p, \sigma_p^2)$ and samples for arm $i$ follow $X_i \sim N(\mu_i, 1)$, where $\mu_i \sim Q_i$, i.e., the mean is sampled from the prior. Further suppose $\mu_p, \sigma_p^2 \to \infty$ s.t. $\sigma_p^2 \to \infty$ and $\frac{\mu_p}{\sigma_p^2} \to 0$.

Thus, we initially draw from the environment $\mu_i \sim Q_i$. Then, the posterior of arm $i$ at time $t$ is given by:

$F_{is}(x) = N(\hat{\mu}_{is}, \frac{1}{s})$, i.e.,

function valued r.v, but with finite dimensional parameterization

$$f_{is}(x) = \sqrt{\frac{s}{2\pi}}\, e^{-(x-\hat{\mu}_{is})^2/(2/s)}$$

$\hat{\mu}_{is}$: empirical estimate of mean at time $t$,

$x \in \mathbb{R}$, $t = 1, 2, \ldots, n$.

pdf → integrate over $(-\infty, z]$ to get $\{F_i(t)(z)$, $F_{is}(z)$

Further,

$$f_{is}(x) = \sqrt{\frac{s}{2\pi}}\, e^{-(x-\hat{\mu}_{is})^2/(2/s)}$$

$\hat{\mu}_{is}$: empirical estimate of mean with $s$ samples

$i$: arm index; $s$: samples; $x$: CDF at $x \in (-\infty, \infty)$

Finally, $G_{is} = 1 - F_{is}(\mu_1 - \varepsilon)$

→ r.v. that is the value of the random function at $(\mu_1 - \varepsilon)$.

Thus, $G_{is} = \int_{\mu_1-\varepsilon}^{\infty} \sqrt{\frac{s}{2\pi}}\, e^{-(x-\hat{\mu}_{is})^2/(2/s)}\, dx$

$\equiv G_{is}(\hat{\mu}_{is})$

$\hat{\mu}_{is}$: r.v. with $\hat{\mu}_{is} \sim N(\mu_i, \frac{1}{s})$

Note: This uses the facts that $X_i$ are iid, $X_i \sim N(\mu_i, 1)$

or equivalently,

$$F_{is}(\mu_1-\varepsilon) = \int_{-\infty}^{\mu_1-\varepsilon} \sqrt{\frac{s}{2\pi}}\, e^{-(x-\hat{\mu}_{is})^2/(2/s)}\, dx.$$

To compute $E\left[\left(\frac{1}{G_{is}} - 1\right)\right]$, we can use this form.

Defining:

$$H(z) = \int_{-\infty}^{z} \sqrt{\frac{s}{2\pi}}\, e^{-y^2/(2/s)}\, dy$$

(CDF of $N(0, 1/s)$ r.v.)

We observe by a (sample-path wise) change of variable: $y = x - \hat{\mu}_{is}$, we have:

$$F_{is}(\mu_1-\varepsilon) = \int_{-\infty}^{\mu_1-\hat{\mu}_{is}-\varepsilon} \sqrt{\frac{s}{2\pi}}\, e^{-y^2/(2/s)}\, dy$$

$$= H(\mu_1 - \hat{\mu}_{is} - \varepsilon) = H(\Delta_i - \varepsilon + R)$$

($\mu_1 - \mu_i$)

$\hat{\mu}_{is} \sim N(\mu_i, \frac{1}{s})$

$R \sim N(0, \frac{1}{s})$.

Note: This uses the facts that $X_i$ are iid, $X_i \sim N(\mu_i, 1)$

Thus, $E\left[\left(\frac{1}{G_{1s}} - 1\right)\right] = E\left[\frac{1 - G_{1s}}{G_{1s}}\right] = E\left[\frac{F_{1s}(\mu_1 - \varepsilon)}{1 - F_{1s}(\mu_1 - \varepsilon)}\right]$

$= E_R\left[\frac{H(R-\varepsilon)}{1 - H(R-\varepsilon)}\right]$ $\quad (\because \Delta_1 = 0)$

$= \int_{-\infty}^{\infty} h(y) \cdot \frac{H(y-\varepsilon)}{1 - H(y-\varepsilon)}\, dy$

$= \int_{-\infty}^{\infty} h(y+\varepsilon) \frac{H(y)}{1 - H(y)}\, dy, \qquad h(y) = \sqrt{\frac{s}{2\pi}}\, e^{-y^2/(2/s)}$

---

**Back to Theorem:**

Note: This proof below is general, i.e., does not require posterior sampling, or any assumptions on prior or dist. of samples. Algorithm model on Page 10 (i.e., sample from $F_i(t)$ and update it based on samples).

Theorem (3b.2 in text): for arm $i$, and $\varepsilon \in \mathbb{R}$,

$$E[T_i(n)] \leq 1 + E\left[\sum_{s=0}^{n-1}\left(\frac{1}{G_{is}} - 1\right)\right] + E\left[\sum_{s=0}^{n-1} \chi_{\{G_{is} > 1/n\}}\right]$$

expectation only over randomness of arm reward and algorithm; NOT over environment

where $G_{is} = 1 - F_{is}(\mu_1 - \varepsilon)$.

**Proof:** $\mathcal{F}_t = \sigma\{A_1, X_1, \ldots, A_t, X_t\}$ ≡ information until (and including) time $t$.

$$W_j(t) = \left\{ \theta_j(t) \leq \mu_1 - \varepsilon \right\}, \quad j = 1, 2, \ldots, k$$

where $\theta_j(t) \sim F_j(t)$ is the sample drawn from the posterior $F_j(t) = F_{j T_j(t-1)}$

↳ simulation sample

# of samples for arm $j$ at the end of time $(t-1)$.

**Note:** We would typically expect (if all estimates are good) that $\theta_j(t) \leq \mu_1 - \varepsilon$, and indeed, with large enough number of samples, $\theta_j(t) \approx \mu_j$.

Further $P\left(\theta_1(t) > \mu_1 - \varepsilon \,\middle|\, \mathcal{F}_{t-1}\right) = 1 - F_{1 T_1(t-1)}(\mu_1 - \varepsilon)$

$= G_{1 T_1(t-1)}$ a.s.

Now, $E[T_j(n)] = E\left[\sum_{t=1}^{n} \chi_{\{A_t = j\}}\right]$

$= E\left[\sum_{t=1}^{n} \chi_{\{A_t = j \cap W_j(t)\}}\right]$ ← $\{\theta_j(t) \leq \mu_1 - \varepsilon\}$

(I)

$+ E\left[\sum_{t=1}^{n} \chi_{\{A_t = j \cap W_j^c(t)\}}\right]$

(II)

(I): Here the $j^{th}$ arm sample is below the true mean of the best arm ($\mu_1$). In this case, we will play arm $j$ (and thus incur regret) if the sample from arm 1, $\theta_1(t)$, is atypical and below it's true value $\mu_1$.

Below, we upper bound (I): $E\left[\sum_{t=1}^{n} \chi_{\{A_t=j \cap W_j(t)\}}\right]$

Let $A_t' = \underset{j\neq 1}{\text{argmax}}\ \theta_j(t)$. Then, we have:

$$P(A_t=1, W_j(t) \mid \mathcal{F}_{t-1}) \geq P(A_t'=j, W_j(t), \theta_1(t) > \mu_1-\varepsilon \mid \mathcal{F}_{t-1})$$

$W_j(t) = \{\theta_j(t) \leq \mu_1-\varepsilon\}$

here $\theta_1(t) > \theta_j(t) \geq \theta_\ell(t)$, $\ell \neq 1$

$\Rightarrow$ this is one way for arm 1 to be the best.

$\{\theta_1(t), \ldots, \theta_k(t)\}$ conditionally indep. given $\mathcal{F}_{t-1}$.

$$= \underbrace{P(\theta_1(t) > \mu_1-\varepsilon \mid \mathcal{F}_{t-1})}_{= G_{1,T_1(t-1)}} \cdot P(\underbrace{A_t'=j, W_j(t)}_{\text{function of } \{\theta_r(t) \ \forall\, r\neq 1\}} \mid \mathcal{F}_{t-1})$$

Now observe: $\longrightarrow$ (I).

$$P(A_t=j, W_j(t) \mid \mathcal{F}_{t-1}) = P(A_t'=j, W_j(t), \theta_1(t) < \theta_j(t) \mid \mathcal{F}_{t-1})$$

($\theta_j(t) \leq \mu_1-\varepsilon$)

best arm = $j$ and $\theta_j(t) \leq \mu_1-\varepsilon$

best arm skipping arm 1 is $j$,

arm 1 below arm $j$

$$\leq P\left(A_t' = j, W_j(t),\ \theta_1(t) \leq \mu_1 - \varepsilon \,\middle|\, \mathcal{F}_{t-1}\right)$$

$$= P\left(A_t' = j, W_j(t) \,\middle|\, \mathcal{F}_{t-1}\right) P\left(\theta_1(t) \leq \mu_1 - \varepsilon \,\middle|\, \mathcal{F}_{t-1}\right)$$

$\{\theta_1(t), \ldots, \theta_k(t)\}$ conditionally indep. given $\mathcal{F}_{t-1}$.

$P\left(\theta_1(t) \leq \mu_1 - \varepsilon \,\middle|\, \mathcal{F}_{t-1}\right) = 1 - G_{1,T_1(t-1)}$

$$\Rightarrow \quad P\left(A_t' = j, W_j(t) \,\middle|\, \mathcal{F}_{t-1}\right) \geq \frac{P\left(A_t = j, W_j(t) \,\middle|\, \mathcal{F}_{t-1}\right)}{1 - G_{1,T_1(t-1)}}$$

$\longrightarrow$ (2).

(1) + (2) $\Rightarrow$

$$P\left(A_t = j, W_j(t) \,\middle|\, \mathcal{F}_{t-1}\right) \leq \left(\frac{1}{G_{1,T_1(t-1)}} - 1\right) P\left(A_t = 1, W_j(t) \,\middle|\, \mathcal{F}_{t-1}\right)$$

$$\leq \left(\frac{1}{G_{1,T_1(t-1)}} - 1\right) P\left(A_t = 1 \,\middle|\, \mathcal{F}_{t-1}\right).$$

$\longrightarrow$ (3).

$$\therefore \quad E\left[\sum_{t=1}^{n} \chi_{\{A_t = j \cap W_j(t)\}}\right]$$

iterated expectations $\rightarrow$

$$= E\left[\sum_{t=1}^{n} E\left[\chi_{\{A_t = j, W_j(t)\}} \,\middle|\, \mathcal{F}_{t-1}\right]\right]$$

$$= E\left[\sum_{t=1}^{n} P\left(A_t=j, W_j(t) \mid \mathcal{F}_{t-1}\right)\right]$$

from ③

$$\leq E\left[\sum_{t=1}^{n}\left(\frac{1}{G_{1,T_1(t-1)}}-1\right) P\left(A_t=1 \mid \mathcal{F}_{t-1}\right)\right]$$

$$P\left(A_t=1 \mid \mathcal{F}_{t-1}\right) \rightarrow E\left[\chi_{\{A_t=1\}} \mid \mathcal{F}_{t-1}\right]$$

$$\leq E\left[\sum_{t=1}^{n}\left(\frac{1}{G_{1,T_1(t-1)}}-1\right) \chi_{\{A_t=1\}}\right]$$

switching from time $t$ to samples $s$

$$\leq E\left[\sum_{s=1}^{n}\left(\frac{1}{G_{1,s}}-1\right)\right]$$

The last step is similar to the trick in the UCB proof. It follows because every time $A_t=1$, $T_1(t-1)$ increments by 1, and thus a term $G_{1,T_1(t-1)}$ appears at most once in the summation for any fixed value of $r=T_1(\cdot)$.

(II): Upper Bounding $E\left[\sum_{t=1}^{n} \chi_{\{A_t=j \cap W_j^c(t)\}}\right]$

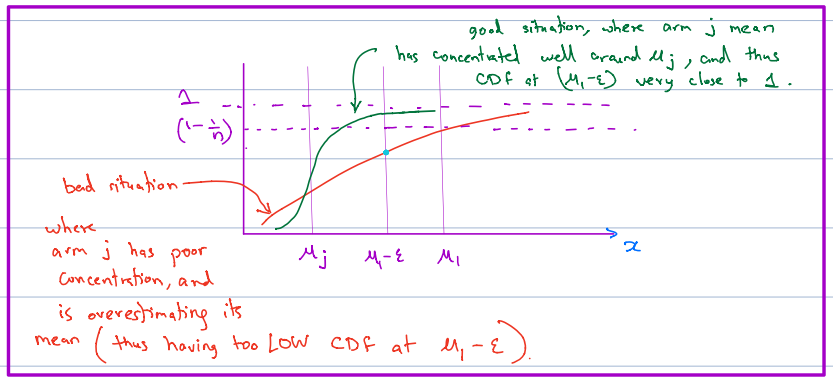
$\{\theta_j(t) > \mu_1-\varepsilon\}$.

This second term controls the number of times that the sub-optimal arm $j$ is played when sample from arm $j$ is atypically high.

Let $\tau = \left\{ t \in \{1, 2, \dots, n\} : 1 - F_{j\, T_j(t-1)}(\mu_1 - \varepsilon) > \frac{1}{n} \right\}$

$$= \left\{ t \in \{1, 2, \dots, n\} : G_{j\, T_j(t-1)} > \frac{1}{n} \right\}$$

These correspond to timeslots s.t. the empirical dist. of arm $j$ has an abnormally large amount of mass above $(\mu_1 - \varepsilon)$ which is well above $\mu_j$, the true mean for arm $j$.




Corresponds to the red empirical CDF in the figure.

Aside: In the proof of TS, we need to control for randomness at two levels: (i) due to the empirical dist. being atypical, and (ii) the sample $\theta_j(t) \sim F_j(t)$ being atypical. The time slots $\tau$ control for (i).

(II): $E\left[\sum_{t=1}^{n} \chi_{\{A_t=j,\, W_j^c(t)\}}\right]$

$\leq E\left[\sum_{t\in\tau} \chi_{\{A_t=j\}}\right]$ → (4).

(III) $+ E\left[\sum_{t\notin\tau} \chi_{\{W_j^c(t)\}}\right]$

(IV)

Now, suppose $A_t=j$, and $T_j(t-1)=(s-1)$, $T_j(t)=s$ for some $t,s$. Then, observe that $G_{j\,T_j(t-1)} > 1/n$, when $t\in\tau$. (Ref: Solutions to Bandit Algorithms, 36.5)

(III): $E\left[\sum_{j\in\tau} \chi_{\{A_t=j\}}\right]$

$\leq E\left[\sum_{t=1}^{n}\sum_{s=1}^{n} \chi_{\left\{T_j(t-1)=s-1,\; T_j(t)=s,\; G_{j\,T_j(t-1)}>\frac{1}{n}\right\}}\right]$

(replace $T_j(t-1)$ by $(s-1)$)

$= E\left[\sum_{s=1}^{n} \chi_{\{G_{j(s-1)}>\frac{1}{n}\}} \underbrace{\sum_{t=1}^{n} \chi_{\{T_j(t-1)=s-1,\; T_j(t)=s\}}}_{\leq 1}\right]$

(for any fixed $s$, this can occur at most once)

$\leq E\left[\sum_{s=1}^{n} \chi_{\{G_{j,(s-1)}>\frac{1}{n}\}}\right]$

(replace $(s-1)$ by $s$ and bound continues to hold.)

(iv) : $E\left[\sum_{t \notin \tau} \chi_{\{W_j^c(t)\}}\right]$ (Ref: Solutions, # 36.5; see textbook website)

$t \notin \tau$ : $\{t : G_{j T_j(t-1)} \leq \frac{1}{n}\}$

$W_j^c(t)$ : $\theta_j(t) > \mu_1 - \varepsilon$

$$= \sum_{t=1}^{n} E\left[E\left[\chi_{\{G_{j T_j(t-1)} \leq \frac{1}{n},\, W_j^c(t)\}} \mid \mathcal{F}_{t-1}\right]\right]$$

$= \chi_{\xi \cdot 3}\, \chi_{\xi \cdot 3}$

$\mathcal{F}_{t-1}$ measurable

---

Now recall: $E\left[\chi_{\{W_j^c(t) \mid \mathcal{F}_{t-1}\}}\right] = P\left(W_j^c(t) \mid \mathcal{F}_{t-1}\right)$

$= P\left(\theta_j(t) > \mu_1 - \varepsilon \mid \mathcal{F}_{t-1}\right) = G_{j T_j(t-1)}$.

---

$t \in \tau$

$$= \sum_{t=1}^{n} E\left[\chi_{\{G_{j T_j(t-1)} \leq \frac{1}{n}\}} \cdot G_{j T_j(t-1)}\right]$$

$\leq 1/n$

$$\leq 1$$

The result now follows by putting these together. ∎

---

Main Theorem: Thompson Sampling has logarithmic regret in the instance dependent frequentist setting. Specifically, (Thm 36.3 in text) in a Gaussian-Gaussian setting,

suppose $F_{js} \sim N\left(\hat{\mu}_j(s), \frac{1}{s}\right)$, $s = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$

(see Note 1 on pp. 13).

Then, $\lim_{n\to\infty} \frac{R_n(\pi, \nu)}{\ell_n(n)} \leq \sum_{\{j:\Delta_j>0\}} \left(\frac{2}{\Delta_j}\right)$, where

$\nu \in \Sigma_N^k(1)$.

Proof: (Adapted from Solutions to Bandit Algorithms, #36.6)

Recall from prev. Thm that: $j \in \{1, 2, \ldots, k\}$, we have

$$E[T_j(n)] \leq 1 + E\left[\sum_{s=0}^{n-1} \left(\frac{1}{G_{1s}} - 1\right)\right] + E\left[\sum_{s=0}^{n-1} \chi_{\{G_{js} > 1/n\}}\right]$$

Further, regret decomposition implies

$$R_n(\pi, \nu) = \sum_{\{j:\Delta_j>0\}} \Delta_j E[T_j(n)]$$

Below, we explicitly derive bounds for the two

terms in $E[T_j(n)]$ in the Gaussian setting.

(I): From Note 1 on pp. 13/14, we have:

$$E\left[\left(\frac{1}{G_{1S}} - 1\right)\right]$$

This step $X_i$ iid, $X_i \sim N(\mu_i, 1)$ from Note 1

$$= \int_{-\infty}^{\infty} h(y+\varepsilon)\frac{H(y)}{1-H(y)}\,dy, \qquad h(y) = \sqrt{\frac{s}{2\pi}}e^{-y^2/(2/s)} \quad (N(0,\tfrac{1}{s}))$$

$$H(y) = \int_{-\infty}^{y} h(z)\,dz.$$

The rest is (intricate) Gaussian calculations.

$$H'(y) \geq \frac{e^{-sy^2/2}}{y\sqrt{s} + \sqrt{sy^2+4}}$$

Fact: $\phi(z)$ CDF of $N(0,1)$

$$1-\phi(z) \geq \frac{e^{-y^2/2}}{y + \sqrt{y^2/4}}$$

Now

$$\int_{-\infty}^{\infty} h(y+\varepsilon)\frac{H(y)}{1-H(y)}\,dy = \int_{0}^{\infty} h(y+\varepsilon)\underbrace{\frac{H(y)}{1-H(y)}}_{\leq 1}\,dy + \int_{-\infty}^{0} h(y+\varepsilon)\underbrace{\frac{H(y)}{1-H(y)}}_{\geq \frac{1}{2}}\,dy$$

$$\leq \int_{0}^{\infty}\frac{h(y+\varepsilon)}{1-H(y)}\,dy + 2\int_{-\infty}^{0} h(y+\varepsilon)H(y)\,dy$$

$$\leq \int_0^{\infty} \frac{h(y+\varepsilon)\, e^{-sy^2/2}}{y\sqrt{s} + \sqrt{sy^2+4}}\, dy + 2\int_{-\infty}^{\infty} h(y+\varepsilon)\cdot H(y)\, dy$$

$$\leq 2e^{-s\varepsilon^2/4} \int_0^{\infty} e^{-sy\varepsilon} (y\sqrt{s}+1)\sqrt{\frac{s}{2\pi}}\, dy + 2e^{-s\varepsilon^2}$$

$$\leq \frac{2(1+\varepsilon\sqrt{s})}{\varepsilon^2 s\sqrt{2\pi}} e^{-s\varepsilon^2/2} + 2e^{-s\varepsilon^2}$$

$$\therefore E\left[\sum_{s=1}^{n}\left(\frac{1}{G_{1s}} - 1\right)\right] \leq \sum_{s=1}^{n}(\cdot) \leq \sum_{s=1}^{\infty}(\cdot)$$

$$\leq \frac{c}{\varepsilon^2}\ln\left(\frac{1}{\varepsilon}\right)$$

(See solutions manual 36.6 (a))

(ii): $$E\left[\sum_{s=0}^{n-1} \chi_{\{G_{js} > 1/n\}}\right] = \sum_{s=1}^{n-1} P\left(G_{js} > \frac{1}{n}\right)$$

Now, suppose

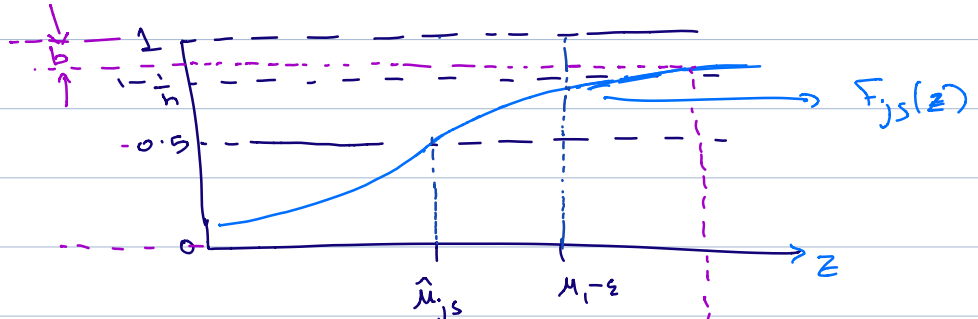
$$G_{js} > \frac{1}{n} \iff 1 - F_{js}(\mu_1 - \varepsilon) > \frac{1}{n}$$

$N\left(\hat{\mu}_{js}, \frac{1}{s}\right)$

$$\iff \quad F_{js}(\mu_1 - \varepsilon) \le \left(1 - \frac{1}{n}\right).$$

(i.e., $\hat{\mu}_{js}$ is "close to" $\mu_1 - \varepsilon$)

↳ wrt $(1/n)$




$\phi^c(x) = 1 - \phi(x)$

↳ std G CDF

$\sigma(\phi^c)^{-1}(b)$

$\sqrt{\frac{1}{s}}$ ↘ ↙ std. Gaussian

Now: $X \sim N(0,1) \Rightarrow P(X > x) \le \frac{e^{-x^2/2}}{\sqrt{2\pi x^2}} \le e^{-x^2/2}$

↳ $x > \frac{1}{\sqrt{2\pi}}$.

$\therefore$ for $x = \sqrt{2\ln(n)}$, $n \ge 3$, we have:

$F_{js}(\mu_1 - \varepsilon) \le 1 - \frac{1}{n} \implies$

$$\hat{\mu}_{js} + \sqrt{\frac{2\ln(n)}{s}} > (\mu_1 - \varepsilon).$$

$$\therefore P\left(G_{js} > \frac{1}{n}\right) \le P\left(\hat{\mu}_{js} + \sqrt{\frac{2\ln(n)}{s}} > \mu_1 - \varepsilon\right)$$

$\mu_1 - \varepsilon$ ↓ $\mu_j + \Delta_j$

$$\leq e^{-\frac{s}{2}\left(\Delta_j - \varepsilon - \sqrt{\frac{2\ln(n)}{s}}\right)^2}$$

$$\therefore \sum_{s=1}^{n} P\left(G_{js} > \frac{1}{n}\right) \leq 1 + \frac{2}{(\Delta_j - \varepsilon)^2}\left(\ln(n) + \sqrt{\pi \ln(n)} + 1\right)$$

(see 36.2 (b), Solutions for more details).

The result now follows by putting the bounds together in the regret decomposition. ∎

Note: The main theorem provides the regret bound as long as the algorithm uses the sampling distribution:

$$F_{is}(x) \sim N\left(\hat{\mu}_{is}, \frac{1}{s}\right),$$

$$i = 1, 2, \ldots, k$$

$$s = 1, 2, \ldots, n.$$

and the rewards $X_i \sim N(\mu_i, 1)$, $i = 1, 2, \ldots, k$.

(a) The regret bound holds <u>without</u> any assumption on the prior, and whether $F_{is}(\cdot)$ above is the correct posterior.

(b) The regret bound matches that of KL-UCB, and thus is a strict improvement over UCB.

(c) In the case when the prior is uninformative, i.e., $Q \sim N(\mu_p, \sigma_p^2)$ with $\sigma_p^2 \to \infty$, $\frac{\mu_p}{\sigma_p^2} \to 0$, then this bound matches the regret lower bound.