Continuous Bayesian Bandits & Open Problems

Yuhua Zhu
University of California - San Diego

Multi-armed Bandits

100 Euro It costs 1 euro to play once $x \sim N(\mu, 1)$, with unknown μ

μ_1 μ_2

Q: What is the best strategy to get the maximum total rewards?

Method 1:

- pull a_1 and a_2 25 times each
- Get the empirical mean $\hat{\mu}_1, \hat{\mu}_2$
- Pull the arm has the larger empirical mean for the rest 50 times

Case 1: $\mu_1 = 10, \quad \mu_2 = -10$

- One does not need to pull 25 times to figure out a_1 is better
- Waste too much time on the obviously worse arm

→ explore too much & exploit too little

Case 2: $\mu_1 = 0.1, \quad \mu_2 = -0.1$

- It is possible that $\hat{\mu}_1 < \hat{\mu}_2$, and the rest 50 times is pulling the worse arm
- Spend too little time to differentiate the two arms

→ explore too little & exploit too much

Tradeoff between "exploitation" and "exploration"

Some better algorithms

Tempered greedy:

Pull a_1 w.p. $\pi_1 \propto e^{c\frac{s_1}{p_1}}$

Pull a_2 w.p. $\pi_2 \propto e^{c\frac{s_2}{q_2}}$

s_i : cumulative reward of a_i

q_i : number of pulls of a_i

UCB:

Compare $\frac{s_1}{q_1} + \sqrt{c\frac{\log(n)}{q_1}}$ & $\frac{s_2}{q_2} + \sqrt{c\frac{\log(n)}{q_2}}$

Multi-armed Bandits

K arms: $\mathscr{A} = \{a_1, \cdots, a_K\}$, Horizon n: the total number of pulls

round i —> pull $A^i \in \mathscr{A}$ —> get reward $X^i \in \mathbb{R}$

policy π^i: $H^i \to \Delta(\mathscr{A})$

Environment $\nu \in \mathscr{E}$: e.q. $x_k \sim N(\nu_k, 1)$, $\nu = (\nu_1, \cdots, \nu_K)$

$\mathscr{E} = \mathbb{R}^K$, each reality <—> $\nu \in \mathscr{E}$

History H^i: $(A^1, X^1, \cdots, A^{i-1}, X^{i-1})$

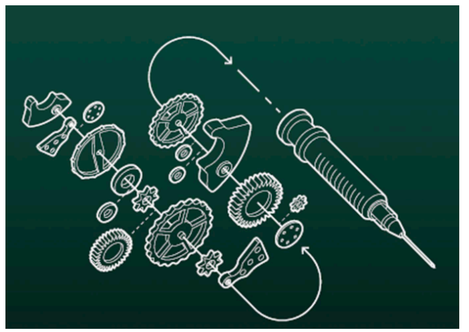
$(s_1^i, q_1^i), \cdots (s_K^i, q_K^i)$ ⟵ Cumulative reward, number of pulls

Goal: find $\{\pi^i\}_{i=1}^n$ to

$$\max \mathbb{E} \sum_{i=1}^{n} X^i$$

Applications

- Advert Placement
- Recommendation Service
- Network Routing
- Dynamic Pricing
- Dynamic treatment



Bayesian v.s. Frequentist

Frequentist:

$$\min R(\nu) = \mathbb{E}\sum_{i=1}^{n}(\mu^*(\nu) - X^i) \longrightarrow \max \mathbb{E}\sum_{i=1}^{n} X^i = \mathbb{E}\sum_{i=1}^{n}\mu_{A^i}(\nu)$$

$$\mu^*(\nu) = \max\{\mu_1(\nu), \cdots, \mu_K(\nu)\},$$ where μ_k is the expectation of arm k

- **It is not an optimization problem.**
- It is proved that a good policy can achieve $R(\nu) = O(log(n))$ for $\forall \nu \in \mathscr{E}$

Bayesian:

$$\max_{\pi} C = \mathbb{E}\sum_{i=1}^{n}\int_{\mathscr{E}} \mu_{A^i}(\nu)\rho^i(\nu)d\nu$$

$\rho^i(\nu)$ is the posterior measure according to the bayesian rule

- **It is an optimization problem**
- The goal is to solve for the best policy that maximizes the expected cumulative reward

History of Bayesian optimal bandits

- [Bradt et al.,1956] developed the framework of bayesian optimal bandits
- Bayesian optimal bandit dominates the research of bandit problem from 1950 - 1980
- The biggest difficulty of bayesian framework is the computational cost:
  e.g. K arms: - exact solve the problem requires $O(n^K)$ computational time
- One breakthrough is the Gitten's index [Gittins, 1979]:
  Computational cost: $O(n^K) \text{ —> } O(Kn^2)$
  However, it can only be applied to infinite discounted cumulative reward:

$$\max C = \mathbb{E} \sum_{i=1}^{\infty} \int_{\mathscr{E}} \gamma^i \mu_{A^i}(\nu) \rho^i(\nu) d\nu$$

- For frequentist setting: [Agrawal, 1995], [Katehakis and Robbins, 1995] developed UCB
  - The computational cost is $O(Kn)$
  - Asymptotically achieve the optimal regret, but not necessarily optimal bayesian rewards

This talk: Continuous Bayesian bandits

Goal: undiscounted bayesian optimal bandit problem

$$\max_{\pi} C = \mathbb{E} \sum_{i=1}^{n} \int_{\mathscr{E}} \mu_{A^i}(\nu) \rho^i(\nu) d\nu$$

- As $n \to \infty$, it converges to a Hamilton-Jacobi-Bellman equation
- Propose a regularized version of bayesian optimal bandits
- Analytic solution exists for some commonly visited cases
  - Computational cost: $O(n^K) \to O(Kn)$
- In general, using numerical solution
  - the computational cost: $O(n^K) \to O(N^K)$ (Depends on mesh)
  - Using DNN to break the curse of dimensionality

An illustrative example

One-armed bandit problem

a_1 gives reward 1 w.p. ν and 0 w.p. $1-\nu$ $\longrightarrow$ $\mu_1(\nu) = \nu$

a_2 gives deterministic reward μ_2

- Initially, before round 1: assume $\nu \sim \text{Beta}(\alpha, \beta)$

- At round i: the posterior measure $\rho^i(\nu)$ depends on
  - s^i: the cumulative reward from a_1 up to round $i-1$
  - q^i: the number of pulls from a_1 up to round $i-1$
  - $\rho^i(\nu) = \rho(\nu \,|\, s^i, q^i) \sim \text{Beta}(\alpha + s^i, \beta + q^i - s^i)$

- a_1 is chosen w.p. $\pi^i = \pi(s^i, q^i)$

- Goal: $\max_{\pi} \mathbb{E}\left[\sum_{i=1}^{n}\left(\int \mu_1(\nu)\rho^i(\nu)d\nu\right)\pi^i + \mu_2(1-\pi^i)\right]$

Discrete bayesian algorithm

- Let $w^i(s,q)$ be the optimal expected cumulative reward starting from round i with $(s^i, q^i) = (s,q)$

- If a_2 is chosen at round i: $w^i(s,q) = w^{i+1}(s,q) + \mu_2$
- If a_1 is chosen at round i:

$$w^i(s,q) = p(s,q) + p(s,q)w^{i+1}(s+1,q+1) + (1-p(s,q))w^{i+1}(s,q+1),$$

  - where $p(s,q) = \dfrac{\alpha + s}{\alpha + \beta + q}$ is the expected reward obtained at round i
  - after pulling a_1,
    - $q^{i+1} = q^i + 1$
    - $\mathbb{P}(s^{i+1} = s^i + 1 \,|\, s^i) = p(s^i, q^i),$
    - $\mathbb{P}(s^{i+1} = s^i \,|\, s^i) = 1 - p(s^i, q^i)\,.$

$$w^i(s,q) = \max\left\{w^{i+1}(s,q) + \mu_2, \quad p(s,q) + p(s,q)w^{i+1}(s+1,q+1) + (1-p(s,q))w^{i+1}(s,q+1)\right\}$$

With $w^{n+1}(s,q) = 0$, one can solve for $w^i(s,q)$ backwards

Derivation to HJB equation

$$w^i(s,q) = \max\left\{w^{i+1}(s,q) + \mu_2,\quad p(s,q) + p(s,q)w^{i+1}(s+1,q+1) + (1-p(s,q))w^{i+1}(s,q+1)\right\} \qquad w^{n+1}(s,q) = 0$$

$$\text{Let } \hat{s} = \frac{1}{n}s, \hat{q} = \frac{1}{n}q \quad \Big| \quad v^i(\hat{s},\hat{q}) = \frac{1}{n}w^i(s,q)$$

$$v^i(\hat{s},\hat{q}) = \max\left\{\frac{1}{n}\mu_2 + v^{i+1}(\hat{s},\hat{q}), \frac{1}{n}\tilde{p}(\hat{s},\hat{q}) + \tilde{p}(\hat{s},\hat{q})v^{i+1}(\hat{s}+\frac{1}{n},\hat{q}+\frac{1}{n}) + (1-\tilde{p}(\hat{s},\hat{q}))v^{i+1}(\hat{s},\hat{q}+\frac{1}{n})\right\}, \quad \text{where } \tilde{p}(s,q) = \frac{n^{-1}\alpha + s}{n^{-1}(\alpha+\beta)+q}.$$

$$\text{By setting } \delta_t = \delta_q = \delta_s = \frac{1}{n}, \quad \Big|$$

$$\frac{v^{i+1}(\hat{s},\hat{q}) - v^i(\hat{s},\hat{q})}{\delta_t} + \max\left\{\mu_2, \tilde{p}(\hat{s},\hat{q}) + \tilde{p}(\hat{s},\hat{q})\frac{v^{i+1}(\hat{s}+\delta_s,\hat{q}+\delta_q) - v^{i+1}(\hat{s},\hat{q}+\delta_q)}{\delta_s} + \frac{v^{i+1}(\hat{s},\hat{q}+\delta_q) - v^{i+1}(\hat{s},\hat{q})}{\delta_q}\right\} = 0.$$

$$\text{As } n \to \infty \quad \Big| \quad \delta_t, \delta_q, \delta_s \to 0$$

$$\partial_t V(t,s,q) + \max\left\{\mu_2, \hat{p}(s,q) + \hat{p}(s,q)\partial_s V(t,s,q) + \partial_q V(t,s,q)\right\} = 0, \quad V(1,s,q) = 0, \qquad \text{where } \hat{p}(s,q) = \lim_{n\to\infty}\tilde{p}(s,q).$$

$$\partial_t V(t,s,q) + \max_{\pi\in[0,1]}\left\{\hat{p}(s,q) + \hat{p}(s,q)\partial_s V(t,s,q) + \partial_q V(t,s,q) - \mu_2\right\}\pi + \mu_2 = 0, \quad V(1,s,q) = 0,$$

General Results for K-armed bayesian bandits

History at round i: $\mathbf{s}^i = (s_k^i)_{k=1}^K, \quad \mathbf{q}^i = (q_k^i)_{k=1}^K$

Posterior measure at round i: $\rho^i(\nu) = \rho(\nu \,|\, \mathbf{s}^i, \mathbf{q}^i)$ Policy at round i: $\pi^i = \pi(\mathbf{s}^i, \mathbf{q}^i)$, where $\sum_k \pi_k(s,q) = 1$

Let $\bar{\mu}_k(s,q) = \int \mu_k(\nu)\rho^i(\nu \,|\, s,q)d\nu, \quad \bar{\sigma}_k^2(s,q) = \int x^2 P_k^\nu \rho^i(\nu \,|\, s,q)\, dx\, d\nu, \quad E_k^p(s,q) = \int x^p P_k^\nu \rho^i(\nu \,|\, s,q)\, dx\, d\nu$

Assume $\hat{t} = \dfrac{i-1}{n}, \quad \hat{q} = \dfrac{1}{n}q, \quad \hat{s} = \dfrac{1}{f(n)}s$

Theorem [Z-Ying-Izzo, 22]

If $\lim_{n\to\infty} \dfrac{n}{f(n)}\bar{\mu}_k(f(n)\hat{s}, n\hat{p}) = \hat{\mu}_k(\hat{s},\hat{p}),\ \lim_{n\to\infty} \dfrac{n}{f(n)^2}\bar{\sigma}_k^2(f(n)\hat{s}, n\hat{p}) = \hat{\sigma}_k^2(\hat{s},\hat{p}),\ \lim_{n\to\infty} \dfrac{n}{f(n)^p}E_k^p(f(n)\hat{s}, n\hat{p}) = 0$

then $V(t,\hat{s},\hat{q}) = V(\dfrac{i-1}{n}, \dfrac{s}{f(n)}, \dfrac{q}{n}) = \dfrac{1}{f(n)}w^i(s,q)$ satisfies the following HJB equation:

$$\partial_t V + \max_{\sum_k \pi_k = 1} \{\hat{\mu}_k \partial_{s_k} V + \partial_{q_k} V + \frac{1}{2}\hat{\sigma}_k^2 \partial_{s_k}^2 V + \hat{\mu}_k\}\pi_k = 0 \qquad \text{with } V(1,\hat{s},\hat{q}) = 0, \text{ for } \forall s, q$$

Continuous HJB and control problem

$$\partial_t V + \max_{\sum_k \pi_k = 1} \{\hat{\mu}_k \partial_{s_k} V + \partial_{q_k} V + \frac{1}{2}\hat{\sigma}_k^2 \partial_{s_k}^2 V + \hat{\mu}_k\}\pi_k = 0, \text{ with } V(1,\hat{s},\hat{q}) = 0, \text{ for } \forall s, q$$

$$\Updownarrow$$

$$V(t,\hat{s},\hat{q}) = \max_{\pi} \mathbb{E} \int_t^1 [\hat{\mu}_k(\hat{s}(\tau),\hat{q}(\tau))\pi_k(\hat{s}(\tau),\hat{q}(\tau))]d\tau$$

$$\max_{\pi} \mathbb{E} \int_0^1 \sum_k \hat{\mu}_k(\hat{\mathbf{s}},\hat{\mathbf{q}})\pi_k(\hat{\mathbf{s}},\hat{\mathbf{q}})\, dt$$

s.t. $d\hat{s}_k(t) = \hat{\mu}_k(\hat{\mathbf{s}},\hat{\mathbf{q}})\pi_k(\hat{\mathbf{s}},\hat{\mathbf{q}})dt + \hat{\sigma}_k(\hat{\mathbf{s}},\hat{\mathbf{q}})\sqrt{\pi_k(\hat{\mathbf{s}},\hat{\mathbf{q}})}dB_t$

$$d\hat{q}_k(t) = \pi_k(\hat{s},\hat{q})dt$$

with $\hat{\mathbf{s}}(0) = \mathbf{0}, \quad \hat{\mathbf{q}}(0) = \mathbf{0}\,.$

Regularized HJB equation

$$\max_{\pi} \quad \mathbb{E}\int_0^1 \sum_k \hat{\mu}_k(\hat{s},\hat{q})\pi_k(\hat{s},\hat{q}) \boxed{-\lambda\pi_k \log(\pi_k)} dt$$

$$\text{s.t.} \quad d\hat{s}_k(t) = \hat{\mu}_k(\hat{s},\hat{q})\pi_k(\hat{s},\hat{q})dt + \hat{\sigma}(\hat{s},\hat{q})\sqrt{\pi_k(\hat{s},\hat{q})}dB_t$$

$$d\hat{q}_k(t) = \pi_k(\hat{s},\hat{q})dt$$

$$\text{with} \quad \hat{s}(0) = 0, \quad \hat{q} = 0$$

$$\updownarrow$$

$$\partial_t V + \lambda \log\left(\sum_k \exp\left[\frac{1}{\lambda}H_k\left(\partial_{s_k}V, \partial_{q_k}V, \partial^2_{s_k}V\right)\right]\right) = 0, \text{ with } V(1,\hat{s},\hat{q}) = 0, \text{ for } \forall s, q$$

$$\pi_k \propto \exp\left[\frac{1}{\lambda}H_k\left(\partial_{s_k}V, \partial_{q_k}V, \partial^2_{s_k}V\right)\right]$$

$$\text{Where } H_k\left(p,g,h\right) = \hat{\mu}_k(s,q)p + g + \frac{1}{2}\hat{\sigma}^2_k(s,q)h + \hat{\mu}_k(s,q)$$

Two special cases

Bernoulli rewards:

a_k gives reward γ w.p. ν_k and $-\gamma$ w.p. $1-\nu_k$, (γ is known)

$\mathscr{E} = [0,1]^K$

Prior of $\nu_k \sim \text{Beta}(\alpha_k, \beta_k)$

For $\gamma = O(n^{-p})$, if $\exists f(n) \le O(n^{1/2-p})$, s.t. $\lim_{n\to\infty} \frac{\gamma(\alpha_k - \beta_k)}{f(n)} = \hat{\alpha}_k$, $\lim_{n\to\infty} \frac{\alpha_k + \beta_k}{n} = \hat{\beta}_k$,

$$\longrightarrow \quad \hat{\mu}_k(s,q) = \frac{s + \hat{\alpha}_k}{q + \hat{\beta}_k}, \quad \hat{\sigma}_k \equiv \frac{n^{1/2}}{f(n)}\gamma.$$

Normal rewards:

The reward of a_k follow $N(\nu_k, \sigma^2)$ with known σ

$\mathscr{E} = \mathbb{R}^K$

Prior of $\nu_k \sim N(\alpha_k, \beta_k^2)$

For $\sigma = O(n^{-p})$, if $\exists f(n) \le O(n^{1/2-p})$, s.t. $\lim_{n\to\infty} \frac{\alpha_k \beta_k^{-2}}{f(n)} = \hat{\alpha}_k$, $\lim_{n\to\infty} \frac{\sigma^2 \beta_k^{-2}}{n} = \hat{\beta}_k$,

$$\longrightarrow \quad \hat{\mu}_k(s,q) = \frac{s + \hat{\alpha}_k}{q + \hat{\beta}_k}, \quad \hat{\sigma}_k \equiv \frac{n^{1/2}}{f(n)}\sigma.$$

Analytic solution for the unregularized HJB

Theorem [Z-Ying-Izzo, 22]

when $\hat{\mu}(s,q) = \dfrac{s+\hat{\alpha}_k}{q+\hat{\beta}_k}$, then the optimal policy for the unregularized HJB is

$$\pi_k^*(s,q) = \begin{cases} 1, & k = \text{argmax}_k \left\{ \mu(s_k, q_k) \right\} \\ 0, & o.w. \end{cases}$$

For finite horizon one-armed bandits problem:

$$\text{When } \mu\left(\frac{s}{f(n)}, \frac{q}{n}\right) > \hat{\mu}_2, \pi = 1 \iff \boxed{\frac{s_i}{q_i} + \frac{f(n)(\hat{\alpha} - \hat{\beta}\hat{\mu}_2)}{q_i} > \mu_2}$$

$$\text{Compare with UCB: } \boxed{\frac{s_i}{q_i} + \sqrt{\frac{c\log(n)}{q_i}} > \mu_2}$$

Analytic solution for the regularized HJB

Theorem [Z-Ying-Izzo, 22]

when $\hat{\mu}_k(s,q) = \dfrac{s+\hat{\alpha}_k}{q+\hat{\beta}_k}$, then the optimal policy for the regularized HJB is

$$\pi_k^*(s,q) \propto e^{\mu_k(s_k,q_k)}$$

For finite horizon one-armed bandits problem:

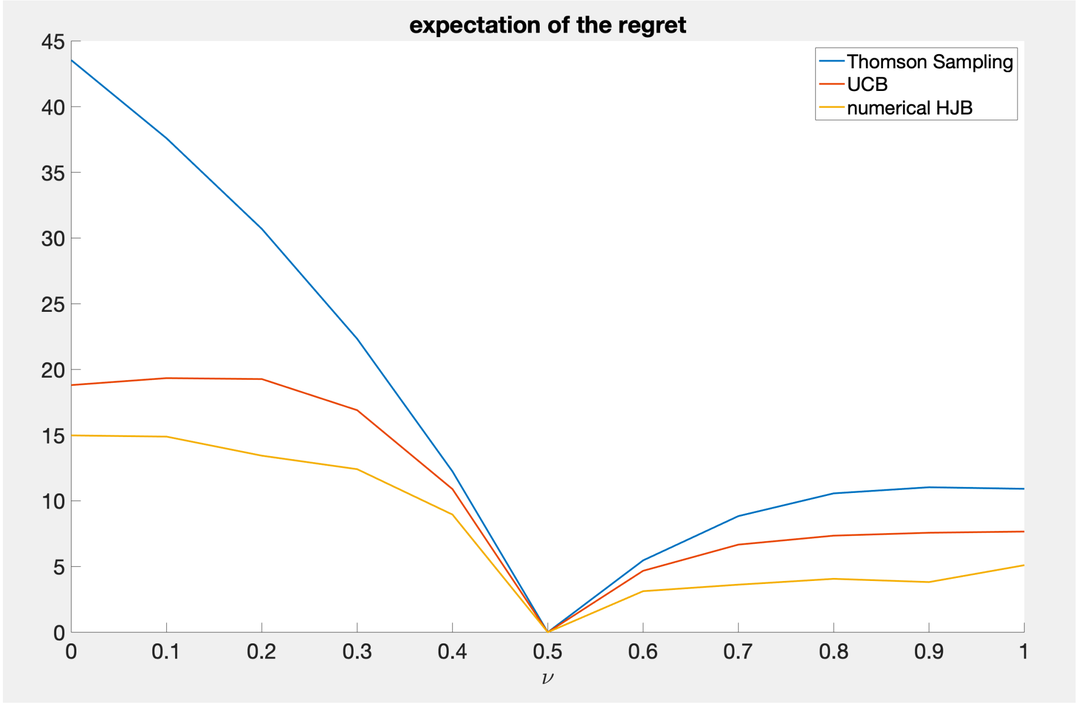
$$\text{When } \mu\left(\frac{s}{f(n)},\frac{q}{n}\right) > \hat{\mu}_2, \pi = 1 \iff \boxed{\pi \propto \exp\left(\frac{n}{f(n)}\frac{s_i+f(n)\hat{\alpha}}{q_i+n\hat{\beta}}\right)}$$

$$\text{Compare with tempered greedy algo: } \boxed{\pi \propto \exp\left(c\frac{s_i+a}{q_i+b}\right)}$$

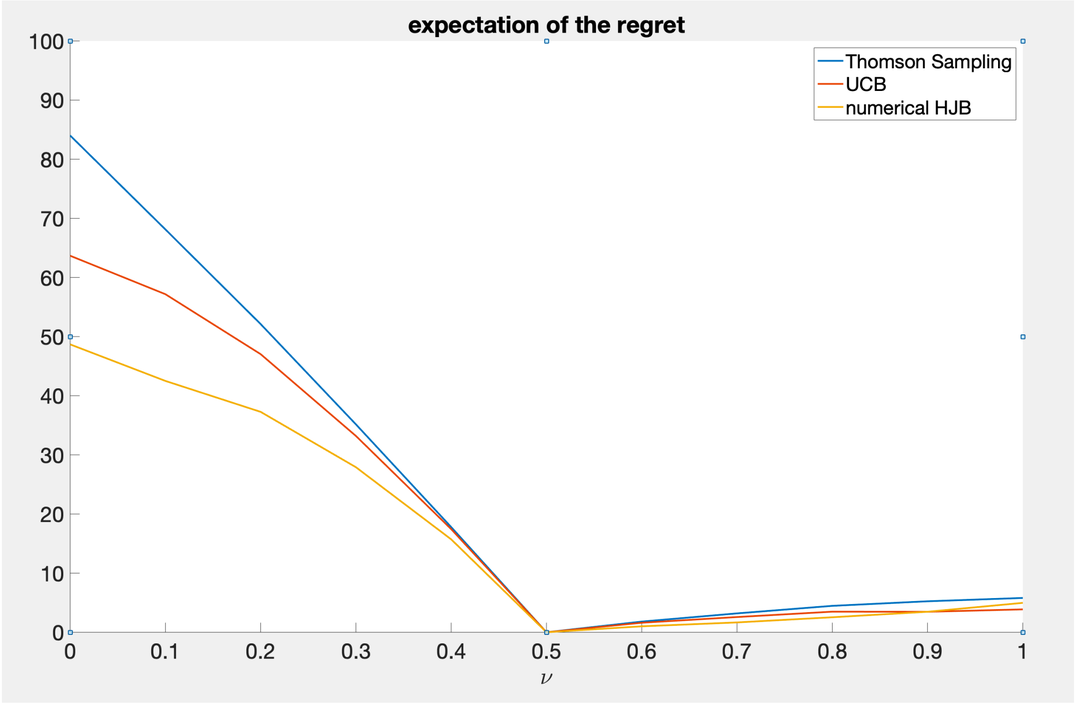
Numerical Experiments

a_k gives reward $\begin{cases} 1 \text{ w.p. } \nu_k; \\ -1 \text{ w.p. } 1-\nu_k \end{cases}$

K = 3, $\nu_1 = 1/2, \nu_2 = \nu_3 = \nu \in [0,1]$



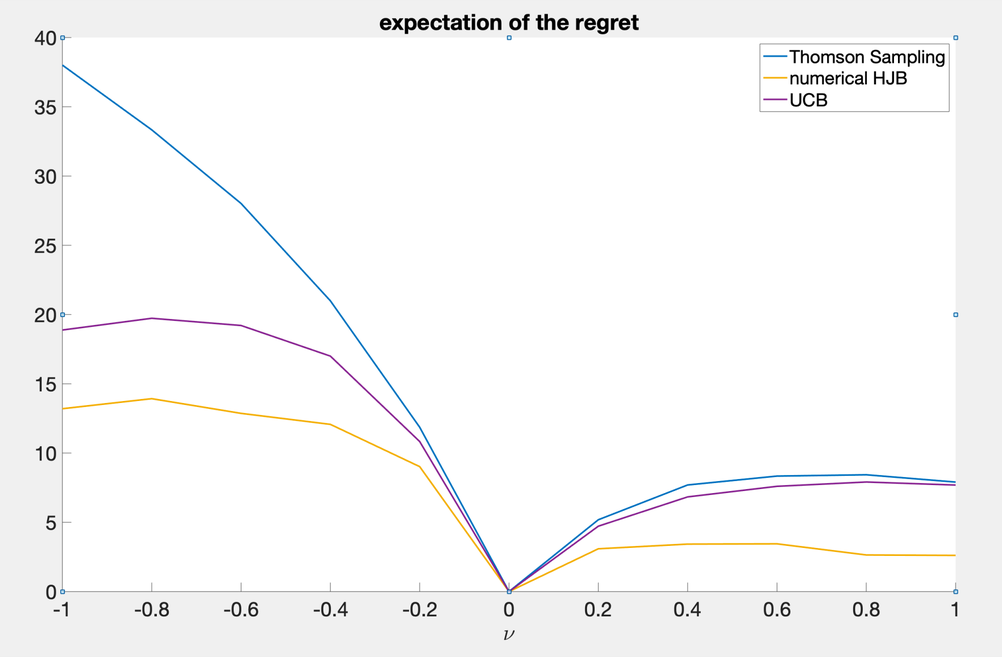
K = 10, $\nu_1 = 1/2, \nu_2 = \cdots = \nu_{10} = \nu \in [0,1]$



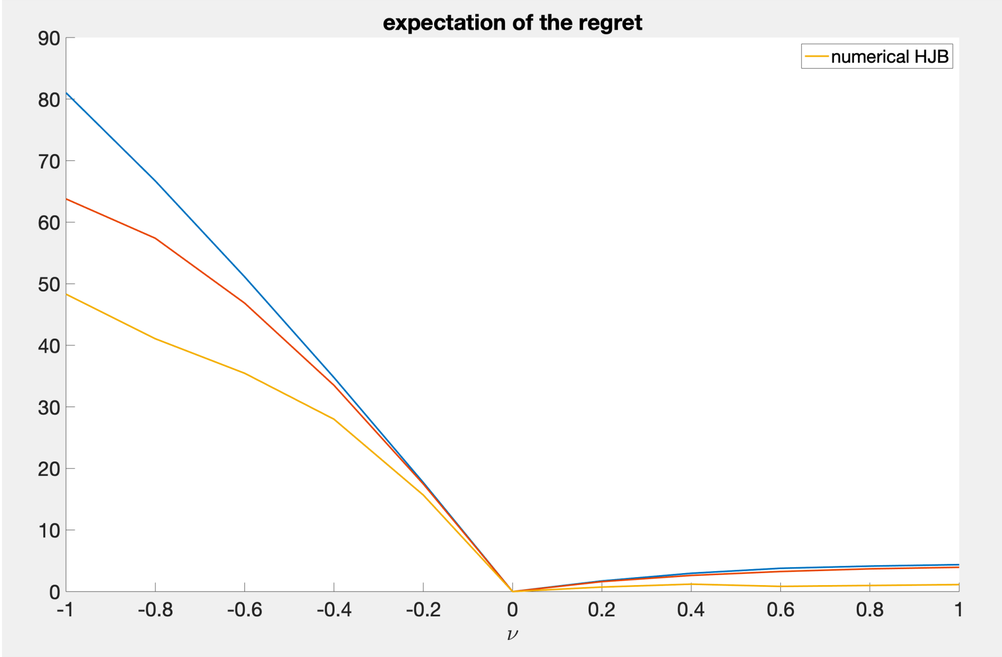
Numerical Experiments

a_k gives reward $\sim N(\nu_k, 1)$

K = 3, $\nu_1 = 0, \nu_2 = \nu_3 = \nu \in [-1,1]$



K = 10, $\nu_1 = 0, \nu_2 = \cdots = \nu_{10} = \nu \in [-1,1]$



Open Problems ?

- When the exact solution is hard to calculate, how to compute the numerical solution efficiently for large K. (DNN?)

- Convergence rate to the HJB equation:
  - $\frac{1}{n} w^i(s,q) \to v(t,\hat{s},\hat{q})$
  - Discrete: $h(n,s,q) > \mu_2 \to \pi = 1$; Continuous: $g(\hat{s},\hat{q}) > \mu_2 \to \pi = 1$
  - $h(n,s,q) \to g\left(\frac{s}{f(n)}, \frac{q}{n}\right)$

- If $\mu(s,q) = \frac{s+\alpha}{q}$, is the HJB equation well-defined?

- Will the HJB equation converge to some “mean-field limit” as $K \to \infty$

- Does there exist a PDE for UCB algorithm as $n \to \infty$?