Approximating the Gittins Index for a Bayesian Bandit

Andrew Lim
National University of Singapore

Joint work with Michael Kim.

Problem statement

- K arms/projects and a retirement option.
- Arms are independent. Expected reward θ_i is constant but unknown; known variance σ_{S_i}, $i = 1, \cdots, K$.
- If arm i is selected in period i, receive reward $Y_i \sim f(\cdot \mid \theta_i)$. If the retirement option is chosen, a one-off endowment M is paid and the process ends.
- Find a scheduling policy that maximizes expected discounted reward over an infinite horizon

Gittins (1979), Whittle (1980)

Dynamic Programming

Let $\rho^i(\theta_i)$ denote the prior on the unknown mean θ_i for arm i and $\mu_i(\rho^i)$ denote the posterior mean.

DP equations for the bandit problem:

$$V(\rho,\, M) = \max\left\{M,\, \max_{i=1,\cdots,K} \mu_i(\rho^i) + \alpha \mathbb{E}_{\rho^i}\left[V(\rho_1^i,\, \rho^{-i}) \,\middle|\, \rho_0 = \rho\right]\right\}$$

Challenging to solve because of the infinite dimensional state space (K-dimensional joint distribution).

Gittins index and Optimal Policy

Consider a one-armed problem with retirement for each arm:

$$V^i(\rho^i, m) = \max\left\{m, \mu_i(\rho^i) + \alpha\mathbb{E}_{\rho^i}V^i(\rho_1^i, m)\right\}$$

GI for arm i is the smallest retirement endowment m such that the decision maker is indifferent between stopping and continuing

$$G^i(\rho^i) = \inf\left\{m : m = \mu(\rho^i) + \alpha\mathbb{E}_{\rho^i}V(\rho_1^i, m)\right\}$$

Optimal policy: If $G^i(\rho^i) \leq M$ for every arm, then retire; else play the arm with the largest Gittins index.

K-armed problem $\rightarrow$ K independent 1-armed bandit problems

Comments

- The Gittins index measures the value of each arm
  - Optimal policy $\equiv$ play the most valuable arm
  - What determines the value of a project?
  - Exploration vs Exploitation
- Gittins index is difficult to compute
  - Solve a fixed point equation that involves the value function of a DP with an infinite dimensional state space.
- Thompson sampling
  - "undiscounted";
  - ignores quality of signal/speed of learning & time horizon

Overview

- Decomposition of the Gittins index into expected value and an "exploration boost"
  - time horizon and quality of the signal

- Approximating the Gittins index
  - Approximation of posterior dynamics
  - Approximate DP $\Rightarrow$ Gittins index
  - Accounts for speed of learning and time horizon

Gittins Index

The smallest retirement endowment M such that the decision maker is indifferent between stopping and continuing

$$G(\rho) = \inf \left\{ M : M = \mu(\rho) + \alpha \mathbb{E}_\rho V(\rho_1,\, M) \right\}$$

This fixed point equation is involves the value function of the one-armed problem

$$V(\rho,\, M) = \max \left\{ M,\, \mu(\rho) + \alpha \mathbb{E}_\rho V(\rho_1,\, M) \right\}$$

Stochastic control problem where the posterior is the state.

Updating the state equations (prior)

Start with the prior $\rho_t(\theta)$.

Given time t observation Y, update the prior via Bayes rule:

$$\rho_{t+1}(\theta) = \frac{\rho_t(\theta) f(Y \mid \theta)}{\int \rho_t(\theta) f(Y \mid \theta) d\theta}$$

The information content of this signal $\sim$ standard deviation σ_S
$\Rightarrow$ determines the speed of learning/amount prior evolves.

Information limits $\sigma_S = 0, \infty$

"No learning" limit ($\sigma_S = \infty$)

- signal has "infinite noise"
- Exploration gives no information about the mean reward θ.

"Perfect learning" limit ($\sigma_S = 0$)

- signal has "zero noise"
- Exploration is very informative and we learn the mean reward after playing once.

Exploration is easy in these limits and should provide lower and upper bounds on the Gittins index.

"No learning" limit ($\sigma_S = \infty$)

- "infinite noise" $\Rightarrow$ posterior does not evolve.

- DP equations

$$V_{NL}(\rho, M) = \max\left\{M,\ \mu(\rho) + \alpha V_{NL}(\rho,\ M)\right\}$$

- Gittins index

$$G_{NL}(\rho) = \frac{\mu(\rho)}{1-\alpha}$$

Value of an arm when there is no learning is its expected reward.
There is no exploration boost since exploration has no value.

"Perfect learning" limit ($\sigma_S = 0$)

- Zero noise in the signal $\Rightarrow$ learn θ after one observation.
- DP equations

$$V_{PL}(\rho) = \max\left\{M,\ \mu(\rho) + \alpha\mathbb{E}_\rho \max\left[M,\ \frac{\theta}{1-\alpha}\right]\right\}$$

- Gittins index:

$$G_{PL}(\rho) = \frac{\mu(\rho)}{1-\alpha} + \alpha\mathbb{E}_\rho \max\left[0,\ \frac{\theta}{1-\alpha} - G_{PL}(\rho)\right]$$

Value of the arm is the expected reward plus a "boost".

Size of the boost depends on the discount factor and the variance of the prior on the mean θ.

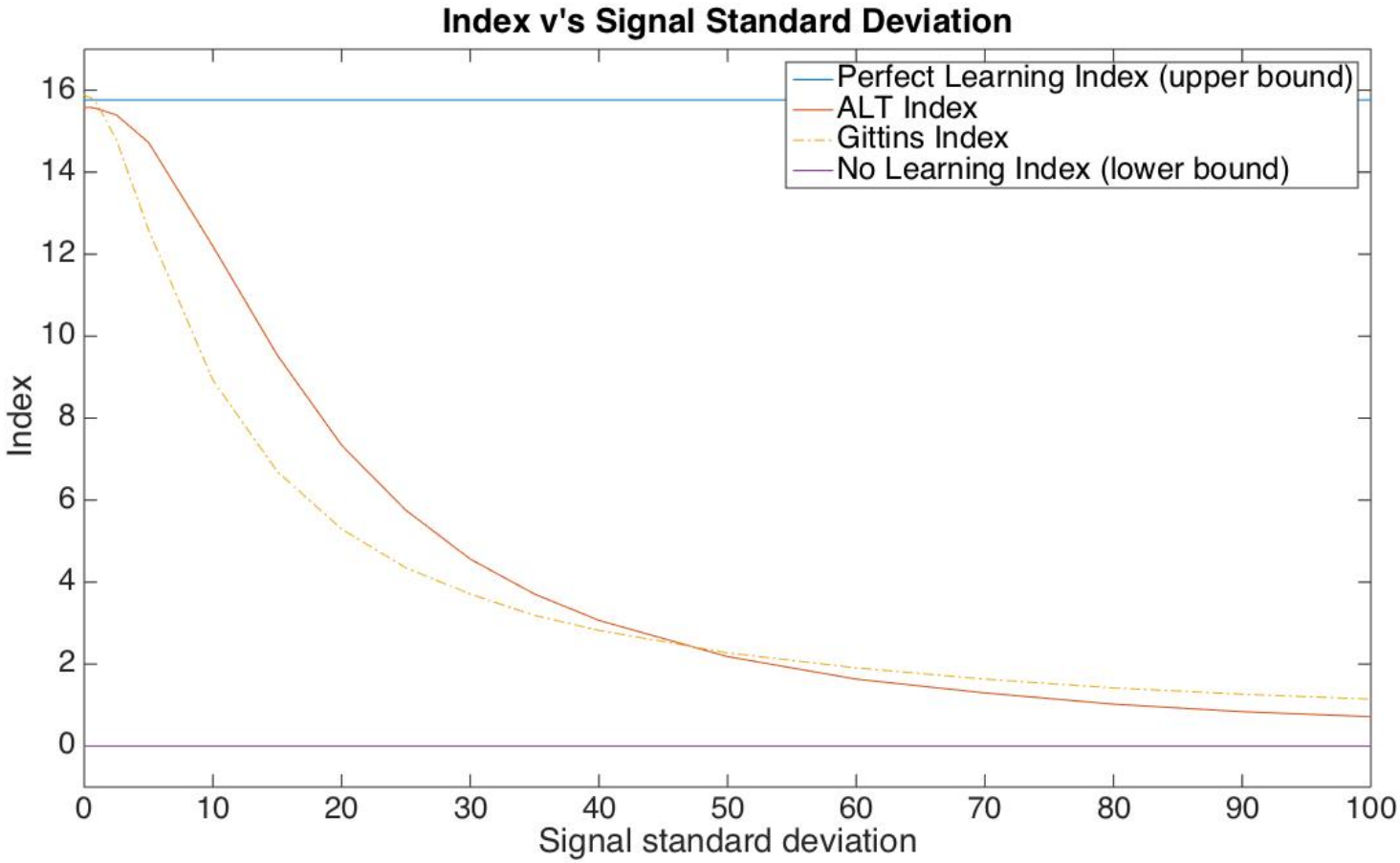
The Gittins index lies between these limits

$$G_{NL} \leq G \leq G_{PL}$$

Difference between these limits captures the impact of learning rate and time horizon on the "value" of an arm

$$G_{PL}(\rho) - G_{NL}(\rho) = \alpha \mathbb{E}_{\rho} \left\{ \max \left[0, \frac{\theta}{1-\alpha} - G_{PL}(\rho) \right] \right\}$$

The difference can be large.


Index v's Signal Standard Deviation
Perfect Learning Index (upper bound)
ALT Index
Gittins Index
No Learning Index (lower bound)
Index
Signal standard deviation
0
2
4
6
8
10
12
14
16
0
10
20
30
40
50
60
70
80
90
100

$$G(\rho) = \frac{\mu(\rho)}{1-\alpha} + \overbrace{\alpha \mathbb{E}_\rho \max\left[0, \frac{\theta}{1-\alpha} - G_{PL}(\rho)\right] - \left[G_{PL}(\rho) - G(\rho)\right]}^{\text{Boost}}$$

The boost is determined by

- the maximum upside potential from PL (2nd term), and
- signal quality (3rd term).

The boost disappears when signals are noninformative ($\sigma_S \to \infty$) or the horizon is short and there is insufficient time to learn sufficiently well and to profit from learning ($\alpha \to 0$)

The first two terms are easy to compute and are independent of signal quality. The 3rd term depends on posterior dynamics and must be computed using DP.

Example: Suppose that

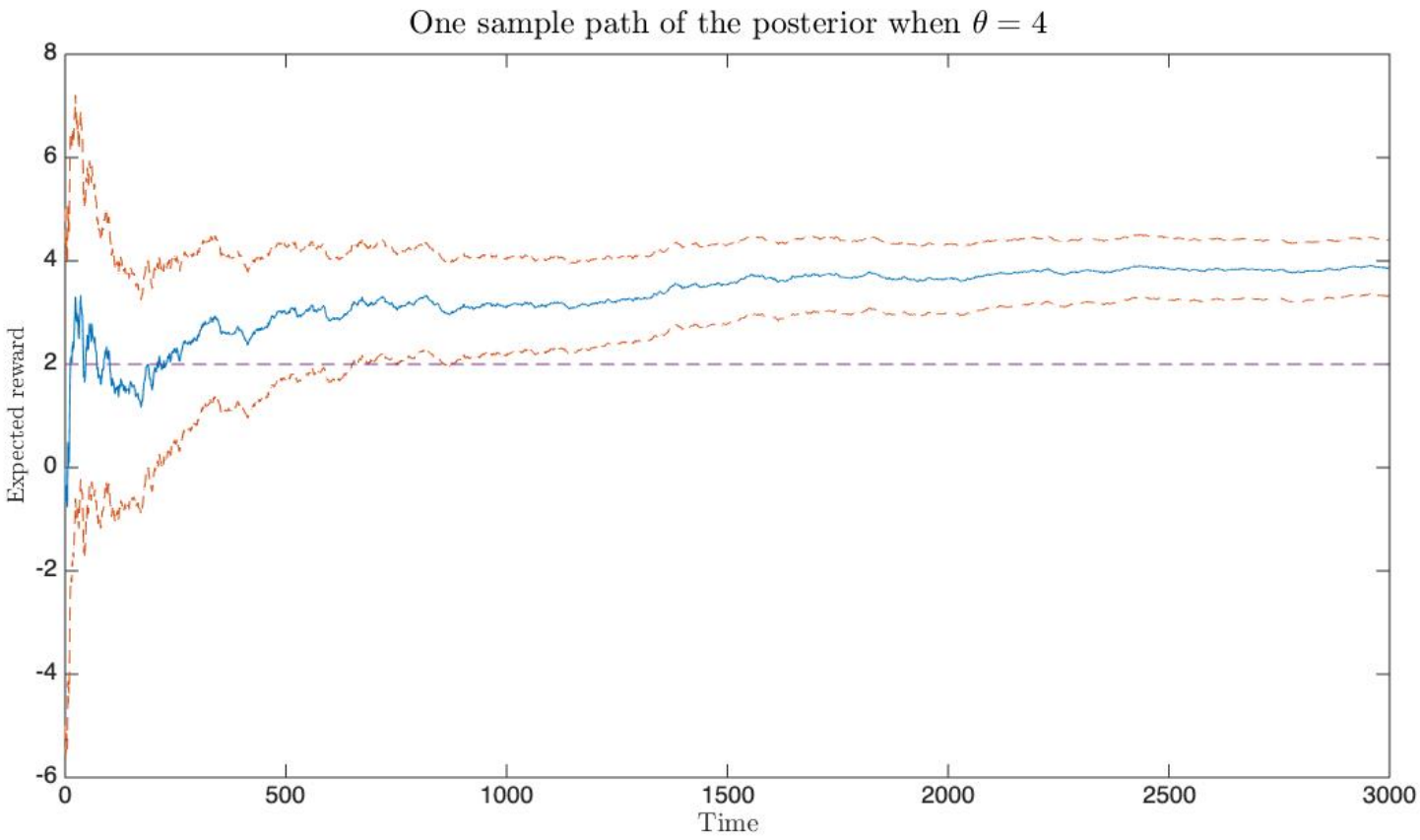
- prior mean = 0; prior sd = 2.5
- signal/reward sd = 15

Expected reward (under the prior) is 0.

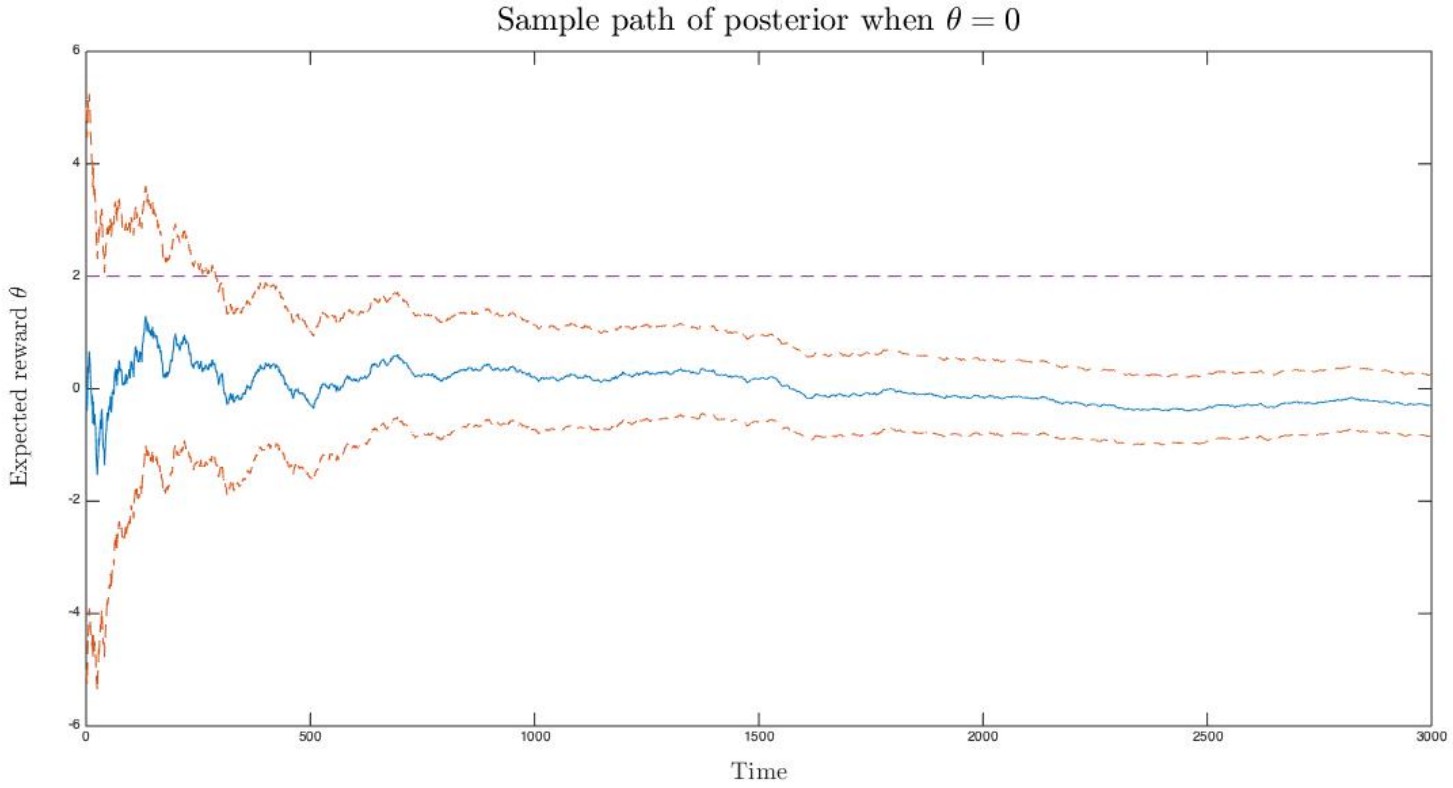
Suppose a retirement endowment equivalent to an annuity of $m = 2$ per year is on the table.

The prior probability of the expected reward exceeding 4 is about 0.05, so $\theta = 4$ is an optimistic scenario.

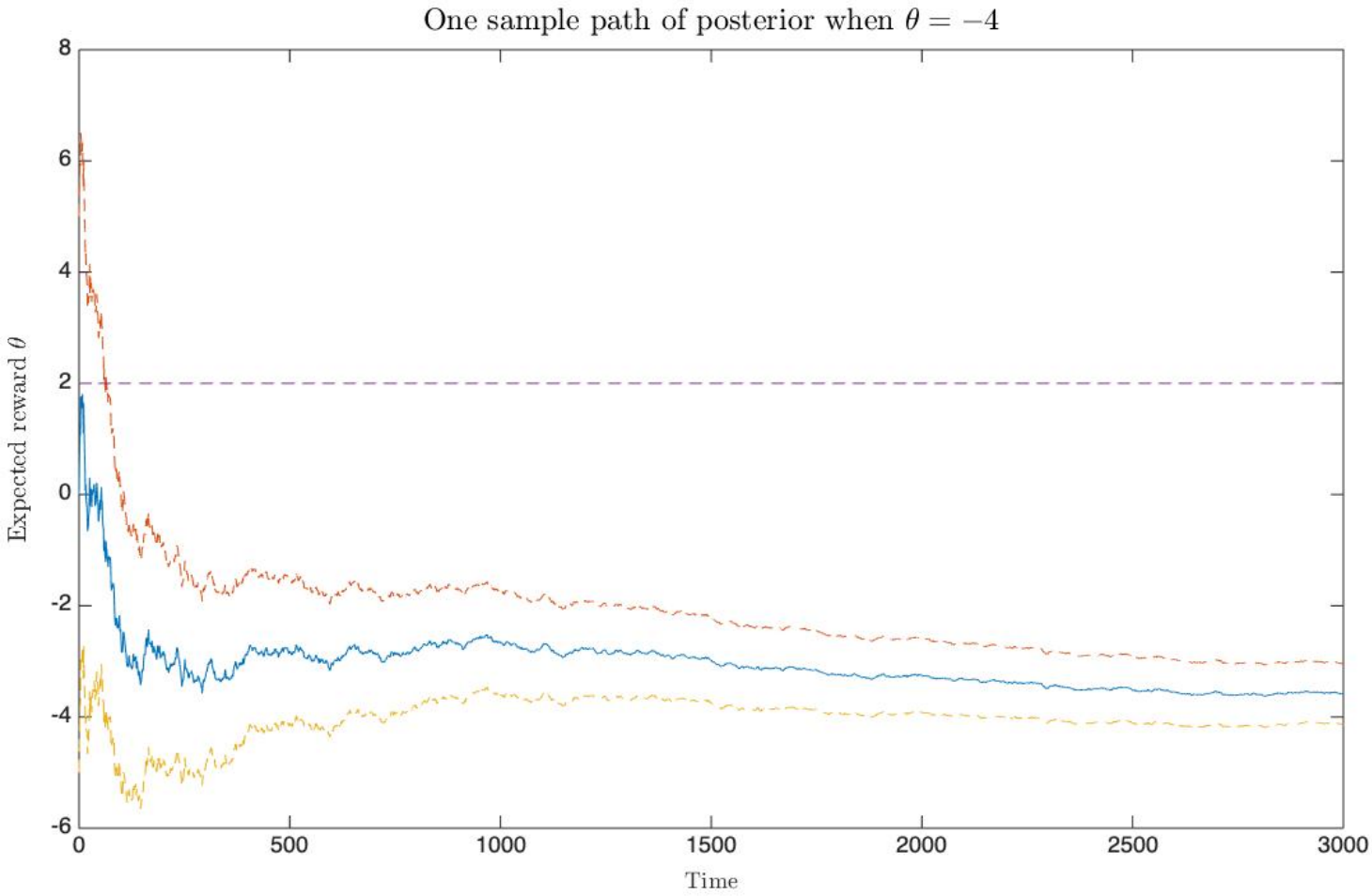
Optimistic case ($\theta = 4$)



Intermediate case ($\theta = 0$)



Worst-case scenario ($\theta = -4$)



Given sufficient time, we can learn how θ compares to the retirement annuity.

e.g. we can learn within about 1000 periods when $m = 2$.

Whether we retire or continue at $T = 0$ depends on the potential upside (e.g. $\theta > 2$ with probability 0.2), the time to profit once we have learned θ, and potential losses while learning
$\Rightarrow$ affects Gittins index (indifference level)

Impact of the discount factor

Discount factor implies a (soft) time horizon.

- time horizon $\sim$ Geometric$(1-\alpha)$
  $\Rightarrow$ expected time horizon $T = 1/(1-\alpha)$.

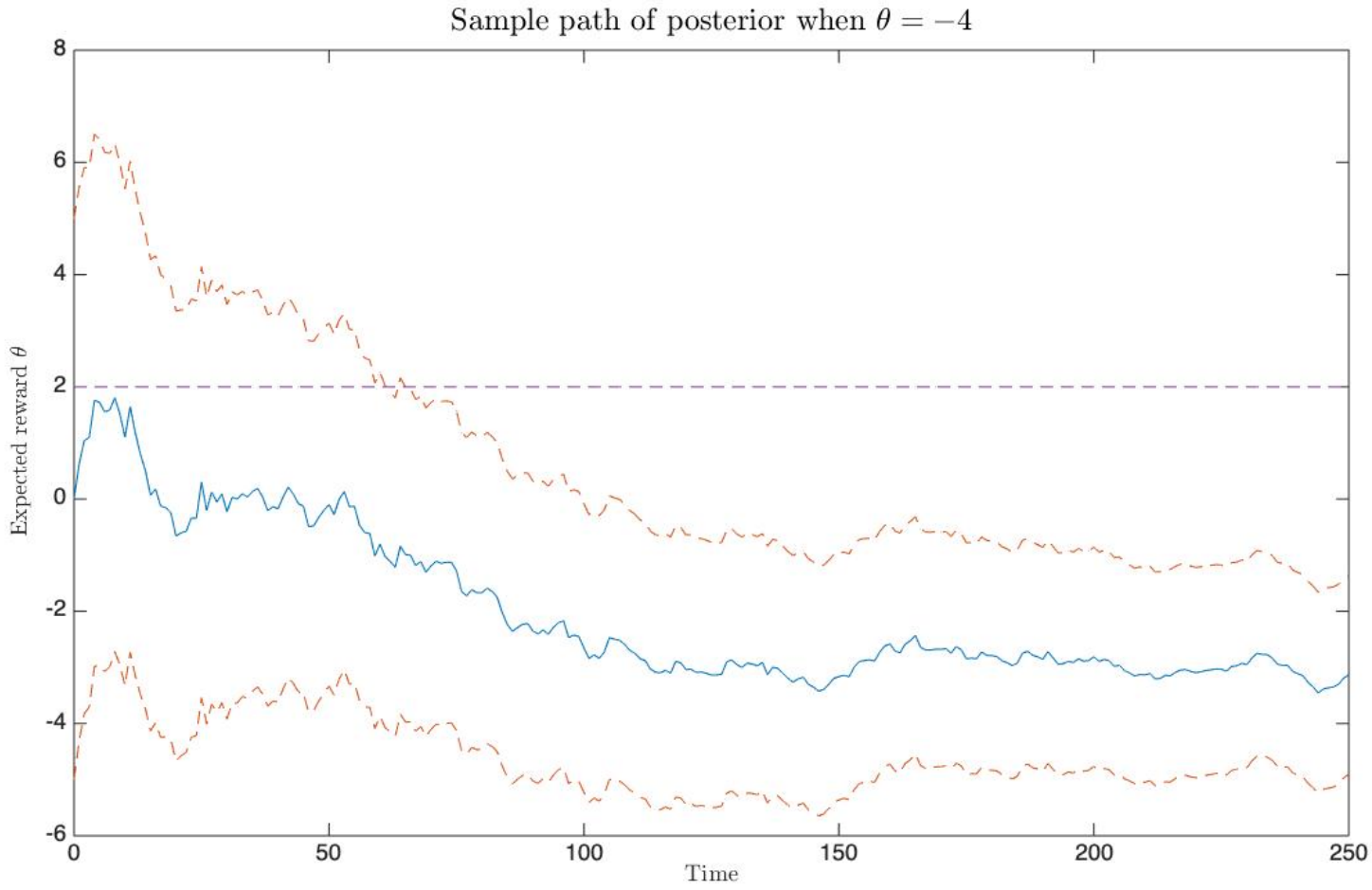
Example:

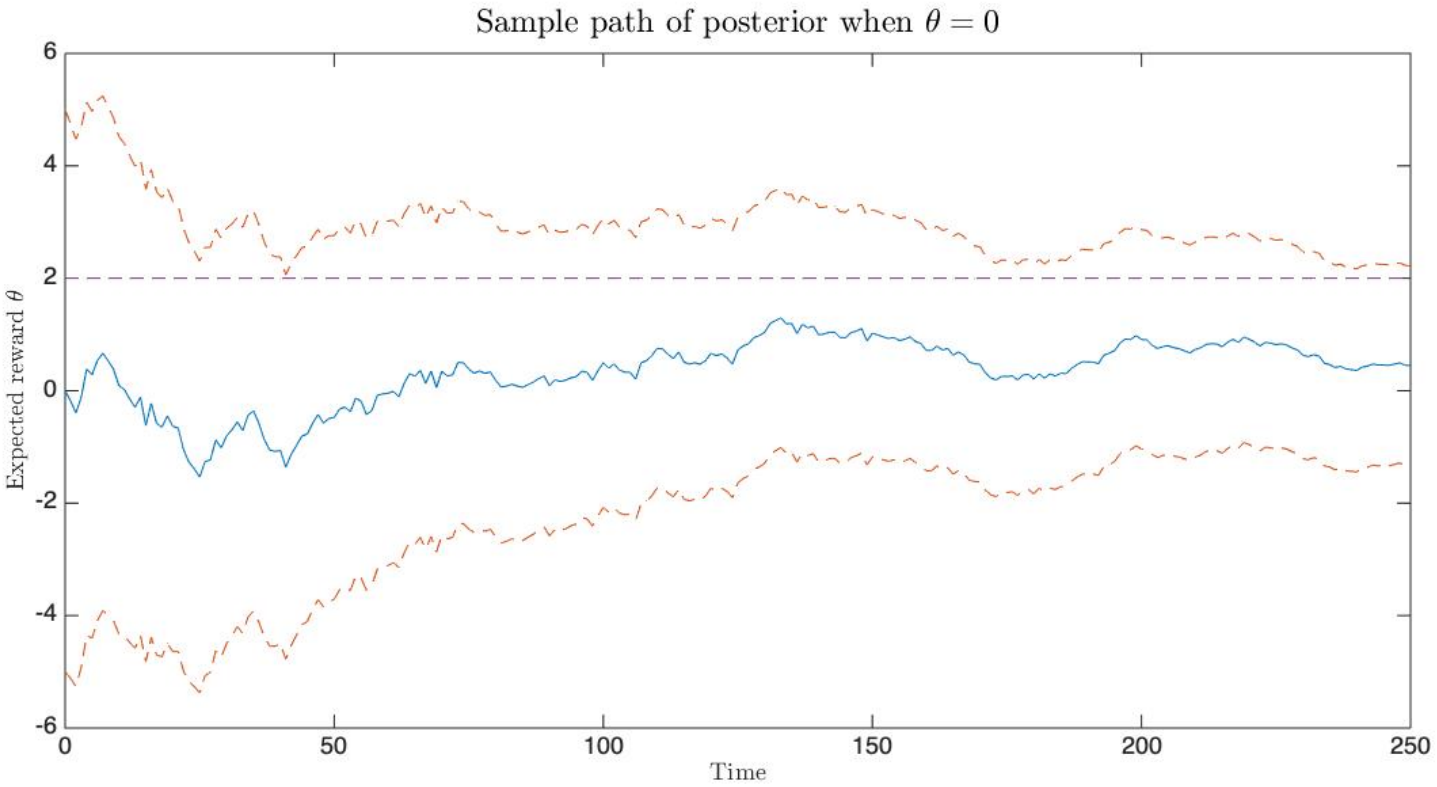
$\alpha = 0.98 \iff T = 1/(1-\alpha) = 50$ years

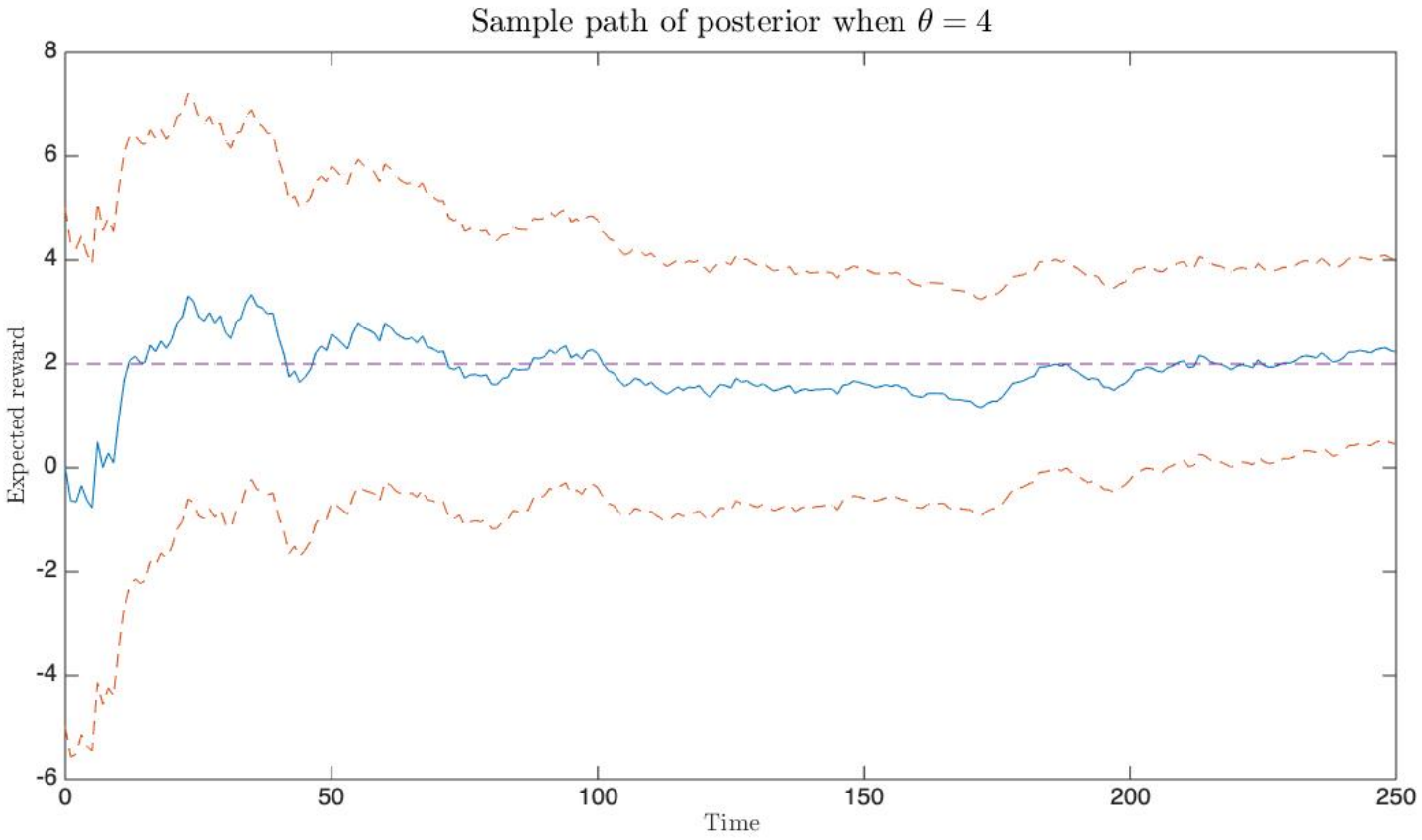
$\alpha = 0.95 \iff T = 1/(1-\alpha) = 20$ years

$\alpha = 0.9 \iff T = 1/(1-\alpha) = 10$ years

We require $\alpha = 0.9997$ for $T = 3000$.


Sample path of posterior when $\theta = -4$
Expected reward θ
Time
8
6
4
2
0
-2
-4
-6
0
50
100
150
200
250


Sample path of posterior when $\theta = 0$
Expected reward θ
Time
6
4
2
0
-2
-4
-6
0
50
100
150
200
250


Sample path of posterior when $\theta = 4$
Expected reward
Time
8
6
4
2
0
-2
-4
-6
0
50
100
150
200
250

- The Gittins index (exploration boost) is determined by the (effective) time horizon and the learning rate.

- Gittins index is difficult to compute because it depends on the value function of a DP with an infinite dimensional state space

Tractable approach to approximately compute the Gittins index.

Approx. of posterior dynamics that accounts for learning rate.

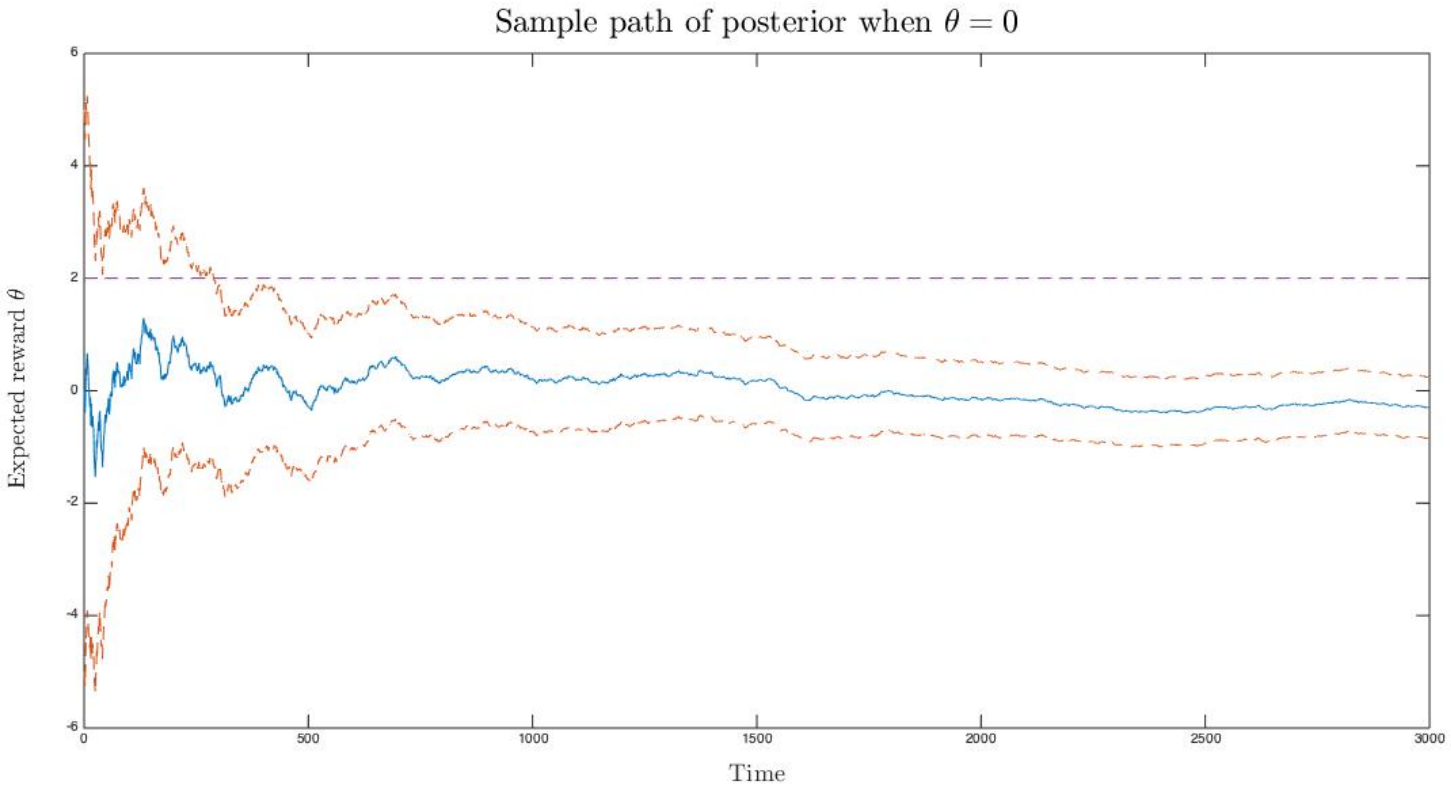
Approximating the Gittins index

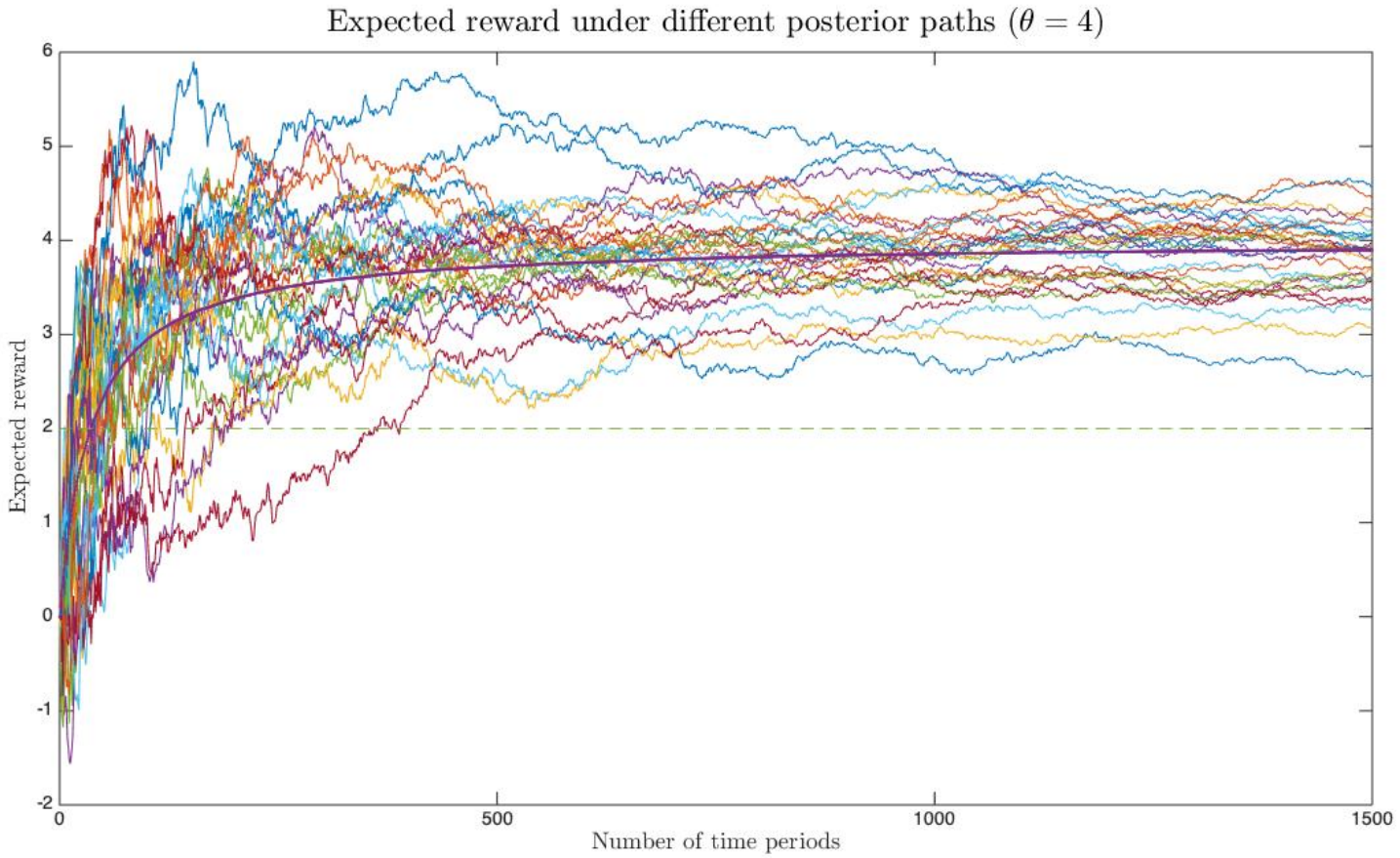
The Gittins index is the solution of a fixed point equation

$$G(\rho) = \inf\left\{M : M = \mu(\rho) + \alpha\mathbb{E}_\rho V(\rho_1,\, M)\right\}$$

where

$$\begin{aligned} V(\rho,\, M) &= \max\left\{M,\, \mu(\rho) + \alpha\mathbb{E}_\rho V(\rho_1,\, M)\right\} \\ &= \sup_{\kappa\geq 0} \mathbb{E}_\rho\left[\sum_{t=0}^{\kappa-1} \alpha^t \mu(\rho_t) + \alpha^\kappa M\right] \end{aligned}$$


Sample path of posterior when $\theta = 0$
Expected reward θ
Time
6
4
2
0
-2
-4
-6
0
500
1000
1500
2000
2500
3000


Expected reward under different posterior paths ($\theta = 4$)
Expected reward
Number of time periods
6
5
4
3
2
1
0
-1
-2
0
500
1000
1500

Bayesian Central Limit Theorem: Given that the mean is θ^*

1. The posterior mean $\mu(\rho_t)$ converges to θ^*

2. The posterior variance is $O(n^{-1})$

3. The posterior is asymptotically normal

$\Rightarrow$ fluid approximation of the conditional posterior dynamics which reduces complexity while still capturing asymptotic consistency and the learning rate.

Posterior

$$\rho_n(\theta) \sim \mathcal{N}\Big(\frac{\tau_P\mu(\rho_0) + n\tau_S\widehat{Y}_n}{n\tau_S + \tau_P},\, n\tau_S + \tau_P\Big)$$

where

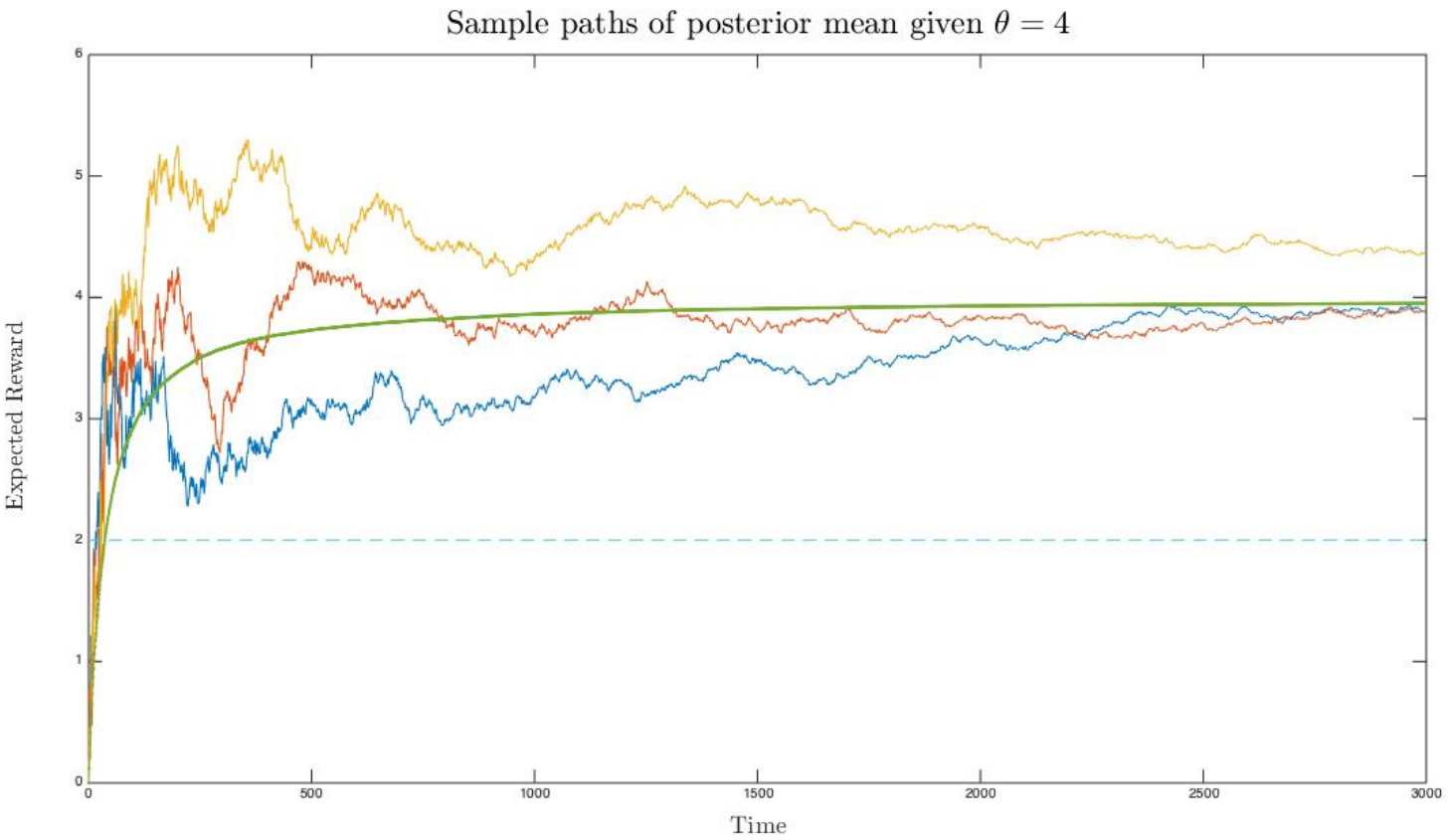
$$\widehat{Y}_n = \frac{Y_1 + \cdots + \widehat{Y}_n}{n}$$

θ-conditional fluid approx. of expected reward under posterior

$$\widehat{\rho}_n(\theta) \sim \mathcal{N}\Big(\frac{\tau_P\mu(\rho_0) + n\tau_S\theta}{n\tau_S + \tau_P},\, n\tau_S + \tau_P\Big)$$

$$\mathbb{E}_{\rho_0}\Big(\mu(\rho_{t+1})\Big|\theta\Big) = \frac{\tau_P\mu(\rho_0) + n\tau_S\theta}{n\tau_S + \tau_P}$$

θ-conditional fluid approximation of the cost-to-go



θ-conditional fluid approximation of the cost-to-go

Approximate the cost-to-go

$$\mathbb{E}_{\rho_0} V(\rho_1, M) \;=\; \mathbb{E}_{\rho_0}\left\{\mathbb{E}_{\rho_0}\Big[\sum_{t=0}^{\kappa^*-1} \alpha^t \mu(\rho_{t+1}) + \alpha^{\kappa^*} M \,\Big|\, \theta\Big]\right\}$$

with a θ-conditional fluid approximation

$$\mathbb{E}_{\rho_0} V^F(\theta, M) = \mathbb{E}_{\rho_0}\left\{\mathbb{E}_{\rho_0}\Big[\sum_{t=0}^{\kappa^*-1} \alpha^t \mathbb{E}_{\rho_0}\Big[\mu(\rho_{t+1}) \,\Big|\, \theta\Big] + \alpha^{\kappa^*} M \,\Big|\, \theta\Big]\right\}$$

Conditional on θ, the (deterministic) trajectory $\mathbb{E}_{\rho_0}\Big[\mu(\rho_{t+1}) \,\Big|\, \theta\Big]$ approximates $\mu(\rho_{t+1})$.

It can be shown that

$$\left|\mathbb{E}_{\rho_0} V(\rho_1,\, M) - \mathbb{E}_{\rho_0} V^F(\theta,\, M)\right| \leq \frac{\mathbb{E}_\tau[\Delta(\tau)]}{1-\alpha}$$

where κ^* is the optimal stopping time associated with $V(\rho_0,\, M)$,

$$\Delta(t) = \mathbb{E}_{\rho_0}\left[\left|\mu(\rho_t) - \mathbb{E}_{\rho_0}\left[\mu(\rho_t)\middle|\theta\right]\right|\right],\ t \geq 1,$$

and τ is a geometric random variable with success probability $1-\alpha$ that is independent of observations.

$\mathbb{E}_\tau[\Delta(\tau)]$ is the expected error in the fluid approximation at the time horizon of the problem.

Gittins index

$$G(\rho) \;=\; \inf\left\{M : M = \mu(\rho) + \alpha\mathbb{E}_\rho V(\rho_1,\, M)\right\}$$

Approximate Gittins index

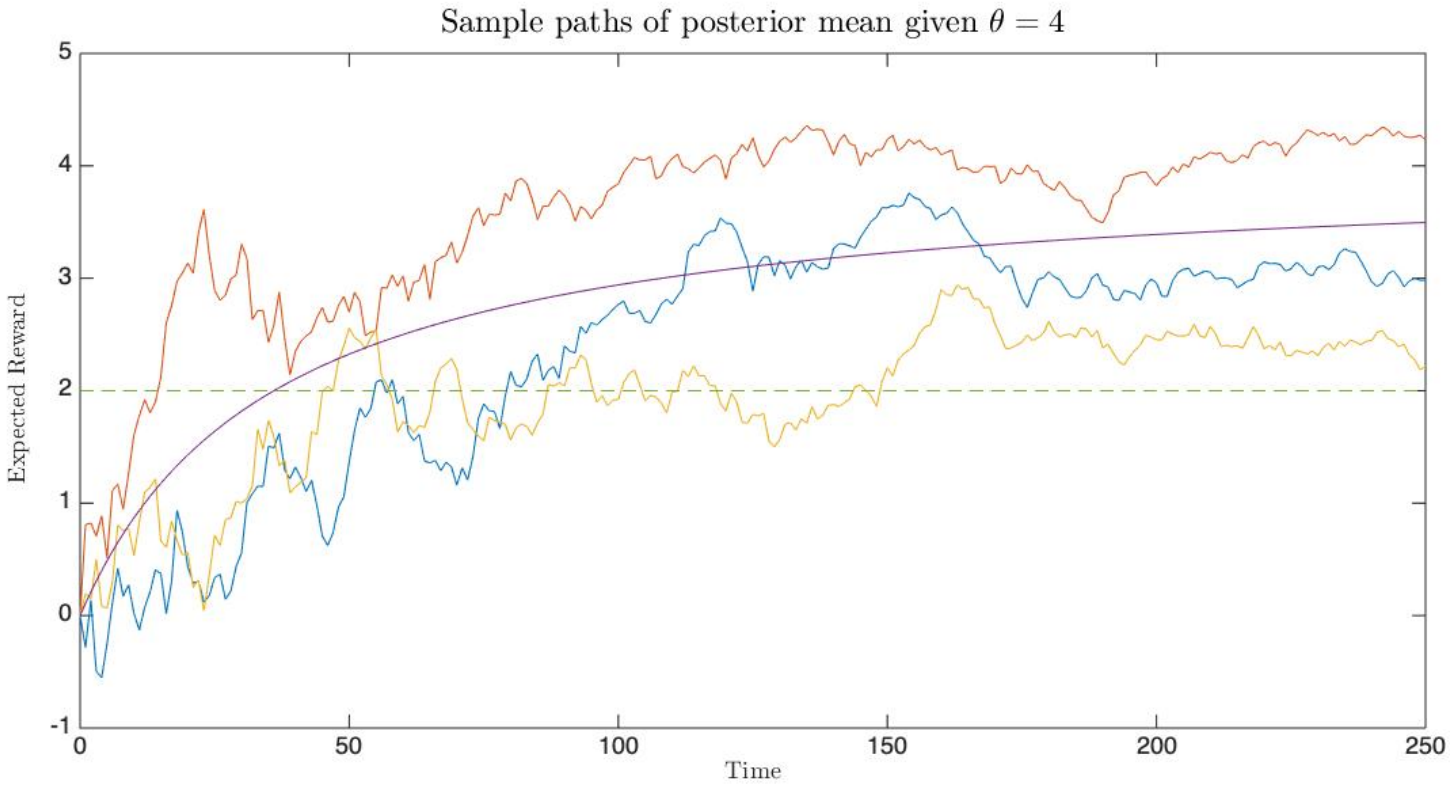
$$G^F(\rho) \;=\; \inf\left\{M : M = \mu(\rho) + \alpha\mathbb{E}_\rho V^F(\theta,\, M)\right\}$$

Bound on approximation error:

$$|G(\rho) - G^F(\rho)| \leq \alpha\frac{\mathbb{E}_\tau[\Delta(\tau)]}{1-\alpha}$$

Unfortunately, $\mathbb{E}_\rho V^F(\theta,\, M)$ is still not easy to compute since it depends on the optimal stopping time of the original problem.

θ-conditional fluid approximation of the cost-to-go



Enlarging the filtration:

$$
\begin{aligned}
V^F(\theta, M) &= \mathbb{E}_{\rho_0}\Big\{ \sum_{t=0}^{\kappa^*-1} \alpha^t \mathbb{E}_{\rho_0}\Big(\mu(\rho_{t+1})\Big|\theta\Big) + \alpha^{\kappa^*} M \,\Big|\, \theta \Big\} \\
&\leq \underbrace{\sup_{\kappa \geq 0} \Big\{ \sum_{t=0}^{\kappa-1} \alpha^t \mathbb{E}_{\rho_0}\Big(\mu(\rho_{t+1})\Big|\theta\Big) + \alpha^{\kappa} M \Big\}}_{J(\theta, M)}
\end{aligned}
$$

e.g. when adopting the normal approximation

$$
\mathbb{E}_{\rho_0}\Big(\mu(\rho_{t+1})\Big|\theta\Big) = \frac{\tau_P \mu(\rho_0) + t\tau_S \theta}{t\tau_S + \tau_P}
$$

Approximate Gittins Index

$$\tilde{G}(\rho) = \inf\left\{M : \ \mu(\rho) + \alpha\mathbb{E}_\rho\left[J(\theta,\, M)\right] = M\right\}$$

$J(\theta,\, M)$ can be computed explicitly

e.g. In the case when $\theta > M(1-\alpha)$ for the normal approximation

$$J(\theta,\, M) = \frac{1}{1-\alpha}\max\left\{M(1-\alpha),\, \mathbb{E}_\tau\left[\mu(\rho)\frac{\tau_P}{\tau_P+\tau_S\tau} + \theta\frac{\tau}{\tau_P+\tau_S\tau}\right]\right\}$$

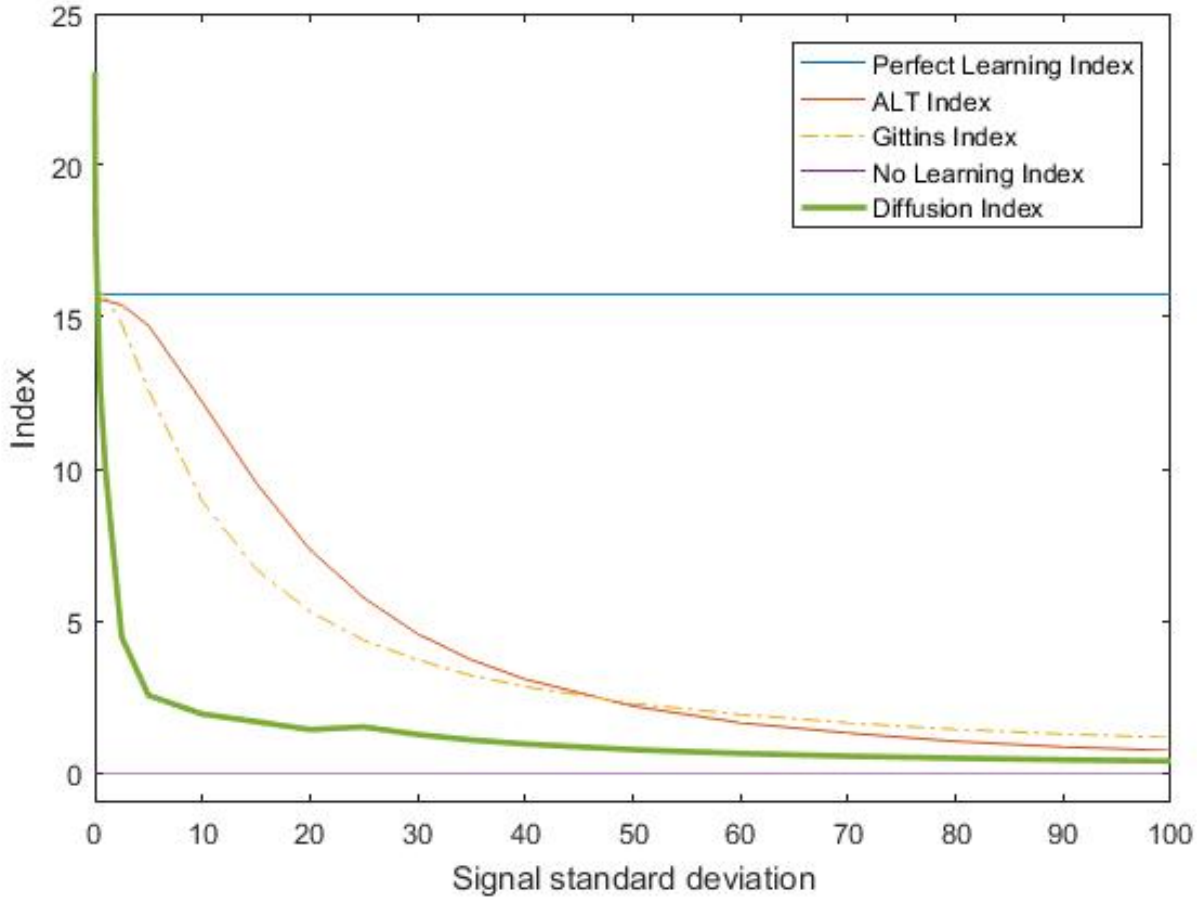
Approximate CTG by sampling θ and approx. the expectation

Sidestep recursion over an infinite-dimensional state space.

Asymptotic optimality

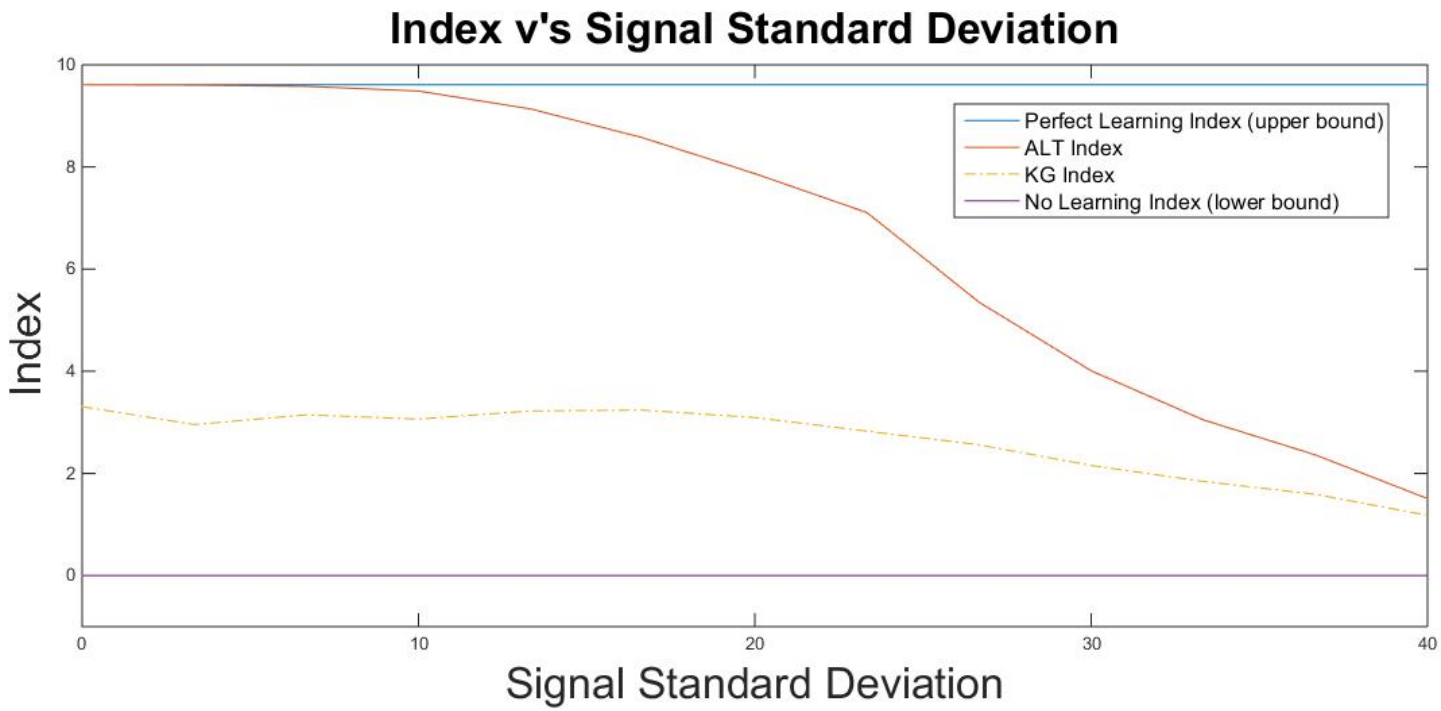
When the signal sd σ_S converges to 0 or ∞, the approximate Gittins index $\tilde{G}(\rho)$ approaches the true Gittins index $G(\rho)$.

The limiting cases also correspond to the Gittins index in the "no learning" and "perfect learning" limits.



Diffusion approximation approach of Brezzi & Lai (2002).

Mixture model



Ryzhov, Powell, Frazier (2012)

Summary

- Gittins index is the value of an arm in a Bayesian bandit
  - Prior, rate of learning, and discount factor (time horizon)
- Tight bounds on the Gittins index that are determined by quality of information
- Approximating posterior dynamics