Online Learning: Applications in Surveillance and Quality Control

James A. Grant

Lancaster University

joint work with D.S. Leslie, K. Glazebrook, R. Szechtman, and A.N. Letchford

Lancaster University

Outline

1. Introduction to online learning problems.
2. Application to surveillance on a perimeter.
3. Application to quality control.

Online Learning

- Traditional optimisation:
  - Typically make one decision
  - If objective function known -> simply* optimise it
  - If objective function uncertain -> estimate expected value -> stochastic optimisation

- Online learning:
  - Initial uncertainty, but opportunity to receive feedback and revise decision
  - Iterate between estimation, decision, and feedback
  - Which decision to make at which stage is non-trivial!

A Simple Example

- Suppose we have an action set of size K, and $T > K$ opportunities to make a decision:
  - One action $k \in \{1, \dots, K\}$ can be chosen at each time $t \in \{1, \dots, T\}$,
  - When chosen, k generates stochastic reward, X_k, with mean μ_k,
  - Aim is to maximise the sum of rewards over T actions.
- If all μ_k known, optimal strategy is to always use $k^* = \mathrm{argmax}_k\, \mu_k$.
- Otherwise, it is necessary to estimate each μ_k.

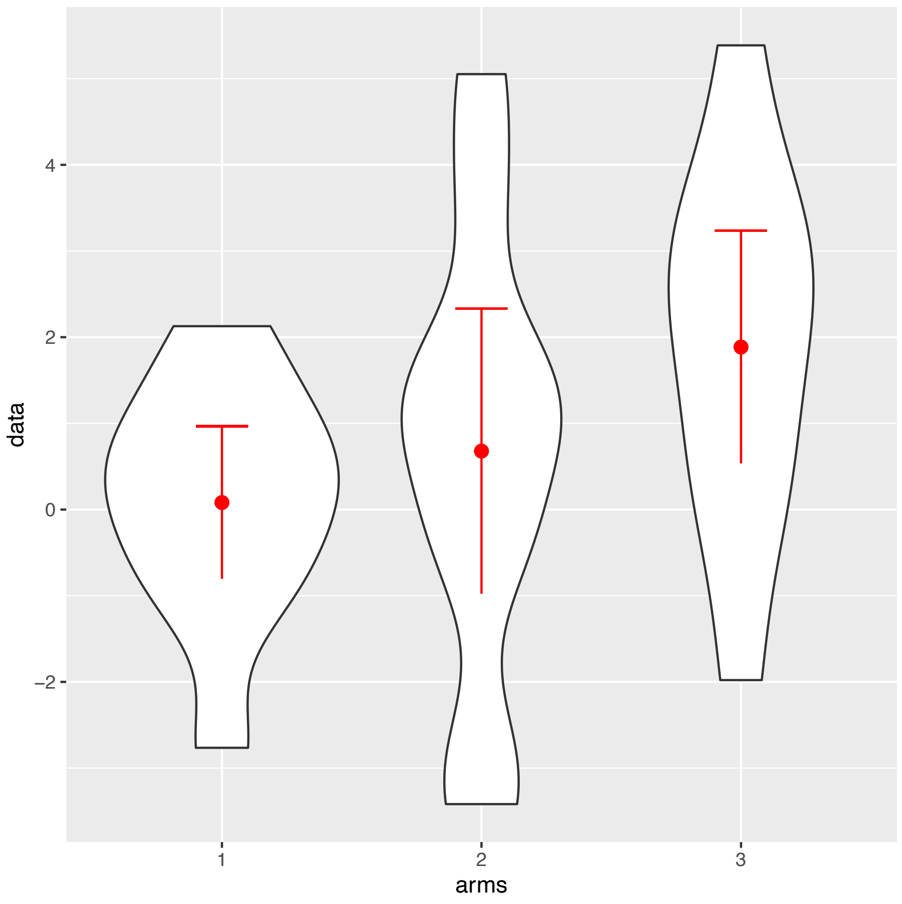
- This problem is known as the *multi-armed bandit problem* – a name derived from a toy application of choosing among K slot machines.

A Naïve Approach

- We could approach this problem by *explore-then-commit:*
  - Use the first $M \cdot K$ rounds to try each action M times,
  - Then compute mean estimators $\hat{\mu}_k = \frac{1}{\mathrm{M}} \sum_m X_{k,m}, \forall k \in \{1, \dots, K\}$,
  - Identify the 'best' action, $k_{max} = \underset{k}{\operatorname{argmax}}\, \hat{\mu}_k$,
  - Use k_{max} at all remaining times $t \in \{MK + 1, \dots, T\}$.

- This will work sometimes, but is sub-optimal in general.
- We need to continue to sample all actions at some level.

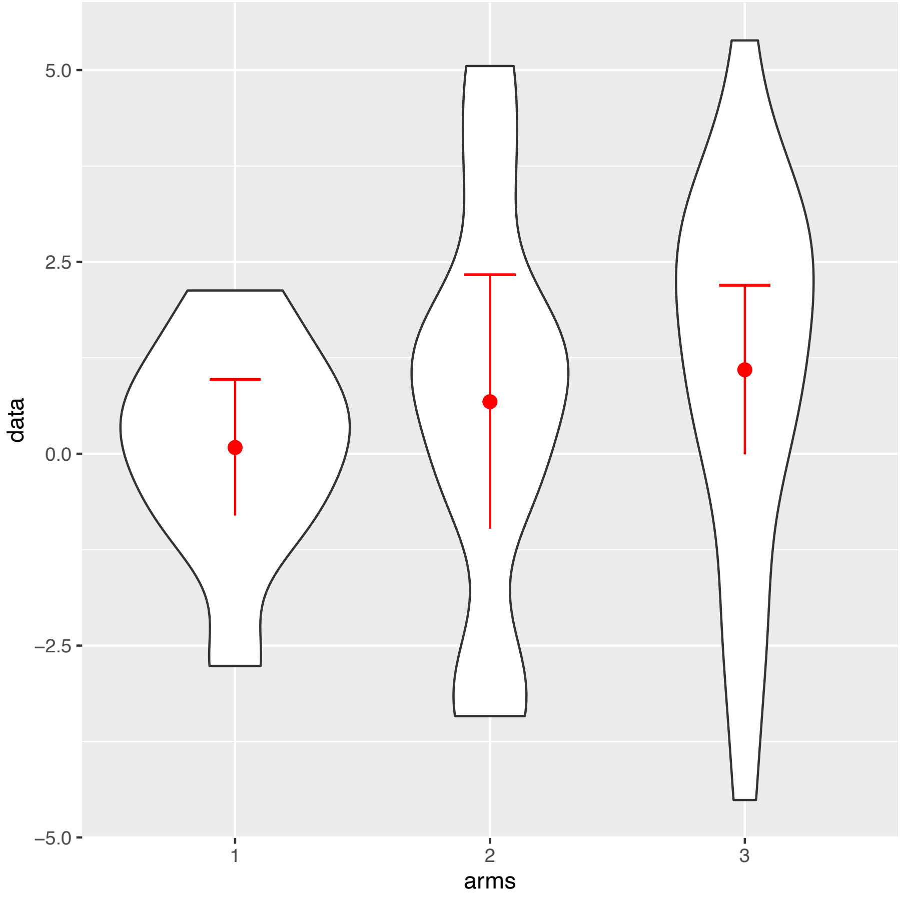
More Successful Strategies



Optimistic Approach

- Consider the upper limit of a confidence interval for each action's mean.

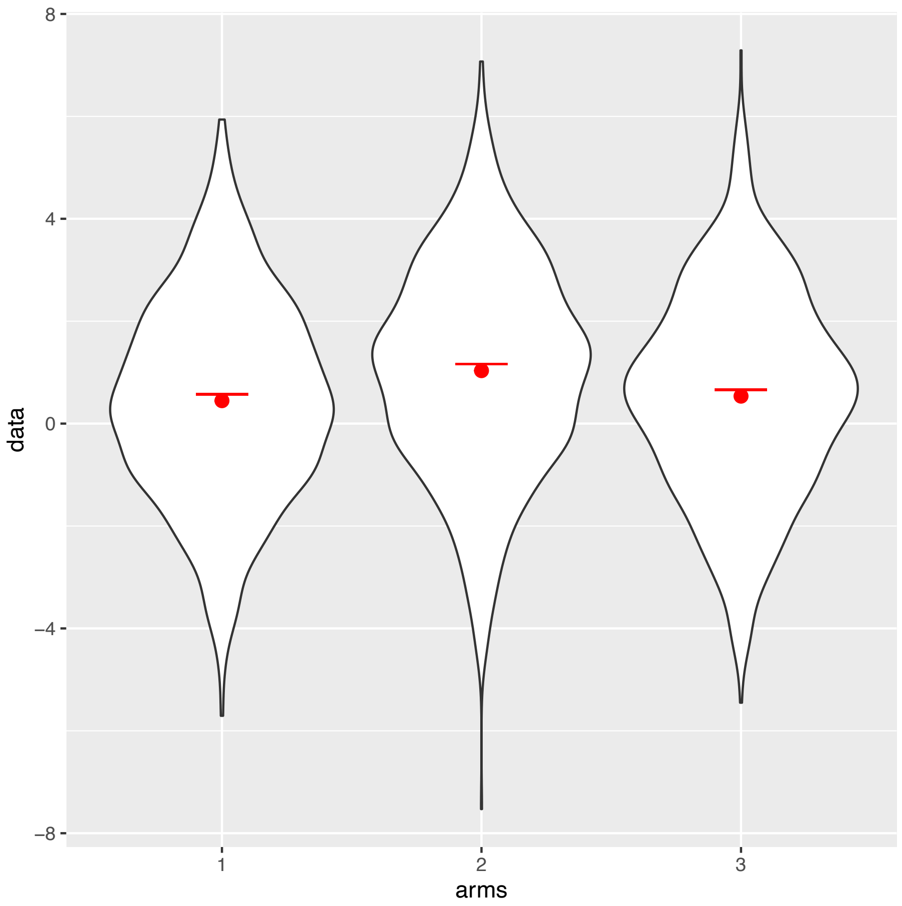
More Successful Strategies



Optimistic Approach

- Consider the upper limit of a confidence interval for each action's mean.
- Deploy the action with the largest upper limit.

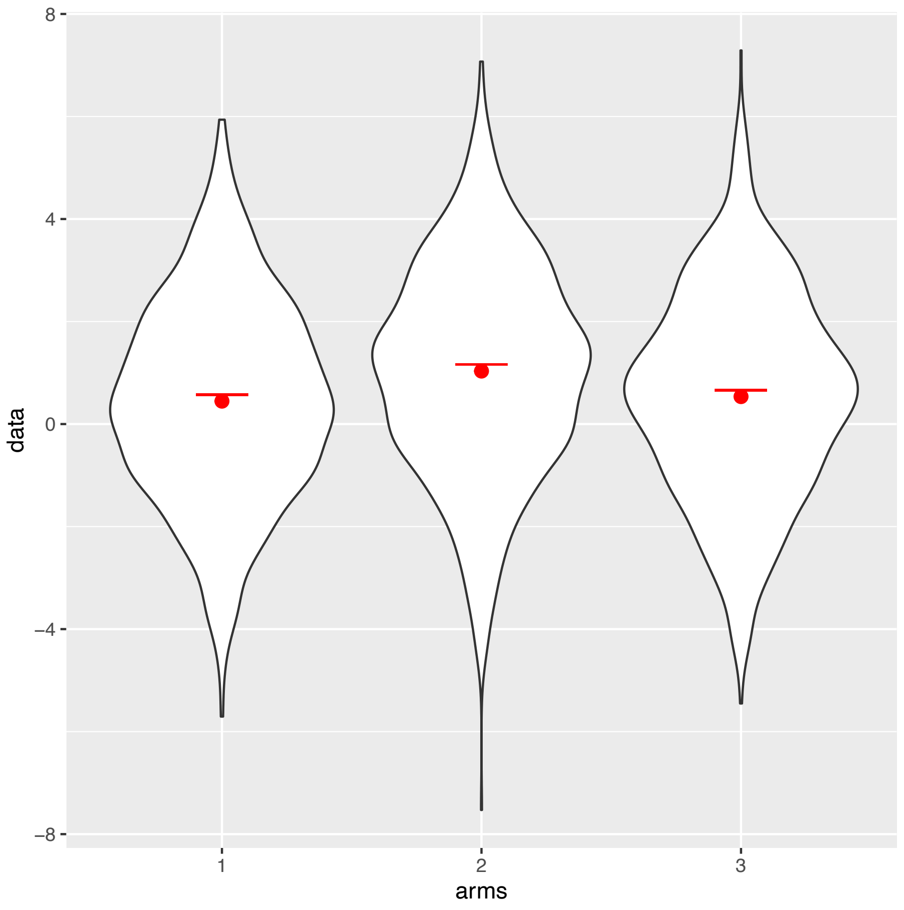
More Successful Strategies



Optimistic Approach

- Consider the upper limit of a confidence interval for each action's mean.
- Deploy the action with the largest upper limit.
- Eventually confidence intervals become small.

More Successful Strategies



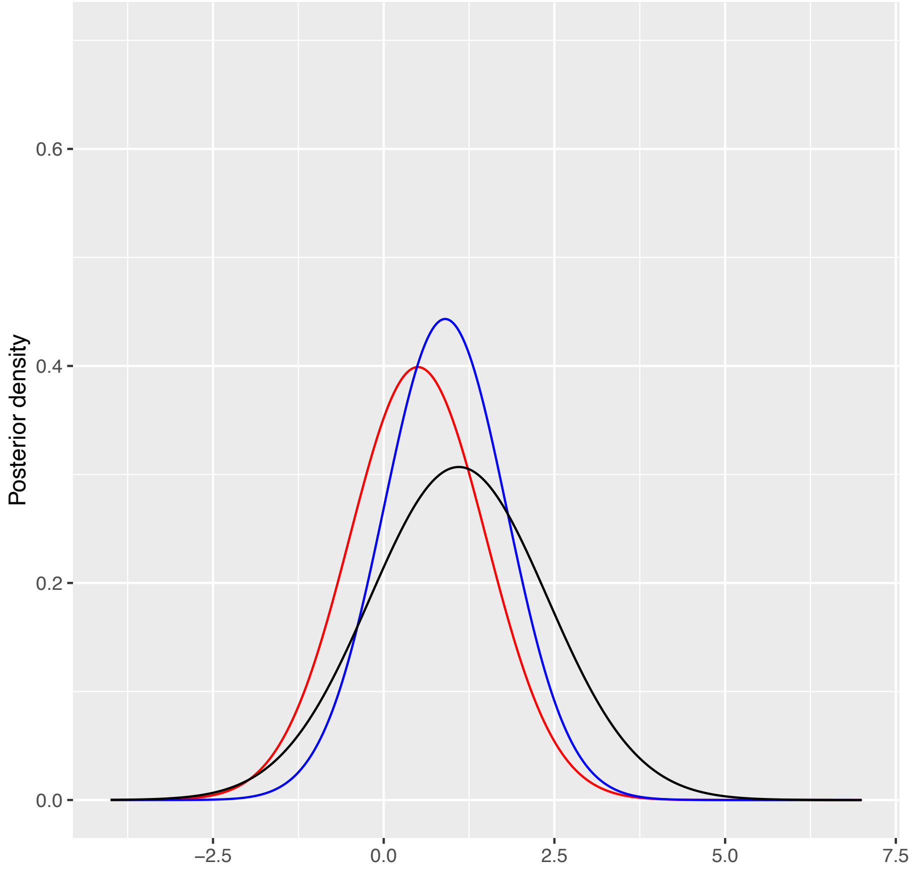
Optimistic Approach

- Consider the upper limit of a confidence interval for each action's mean.
- Deploy the action with the largest upper limit.
- Eventually confidence intervals become small.
- *NB we take increasing quantiles on the limit to ensure exploration.*

More Successful Strategies

Randomised Approach

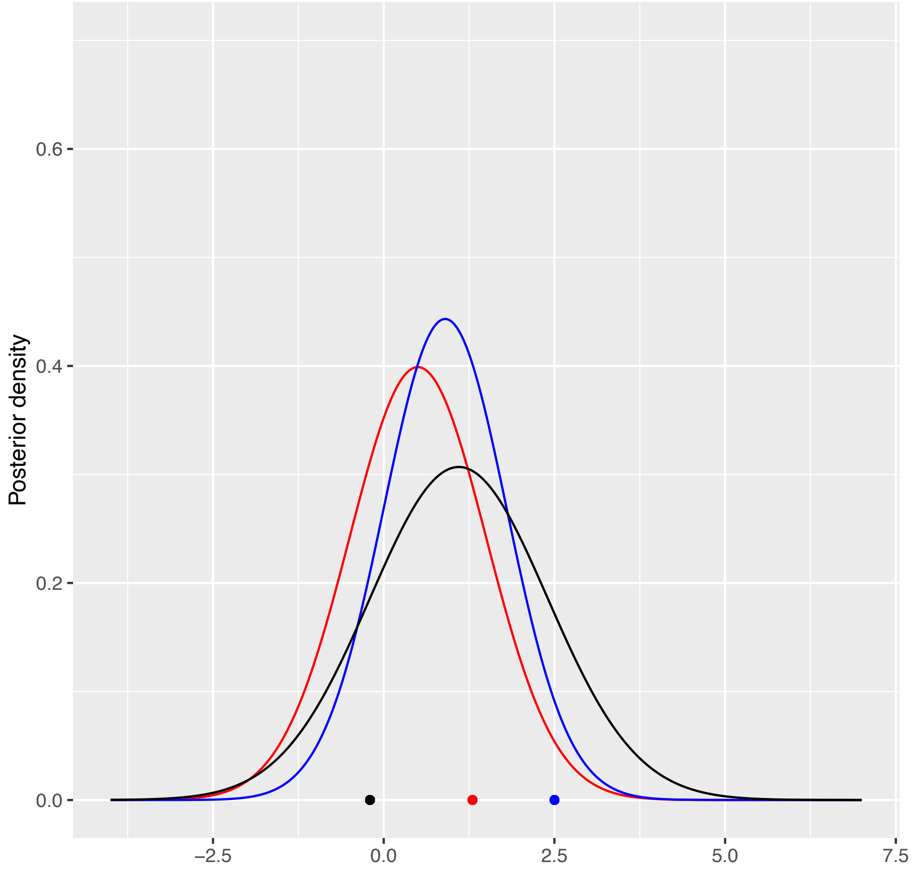
- Consider the posterior distribution on the mean of each action.



More Successful Strategies

Randomised Approach

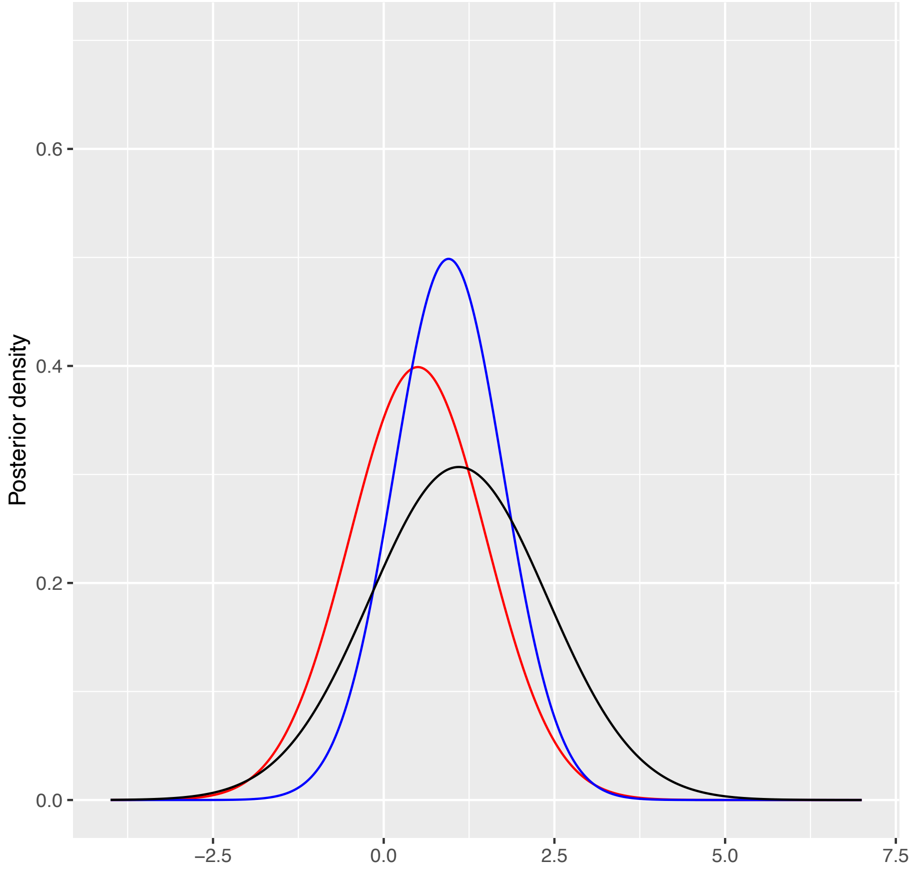
- Consider the posterior distribution on the mean of each action.
- Draw a sample from each and deploy action with the highest sample.



More Successful Strategies

Randomised Approach

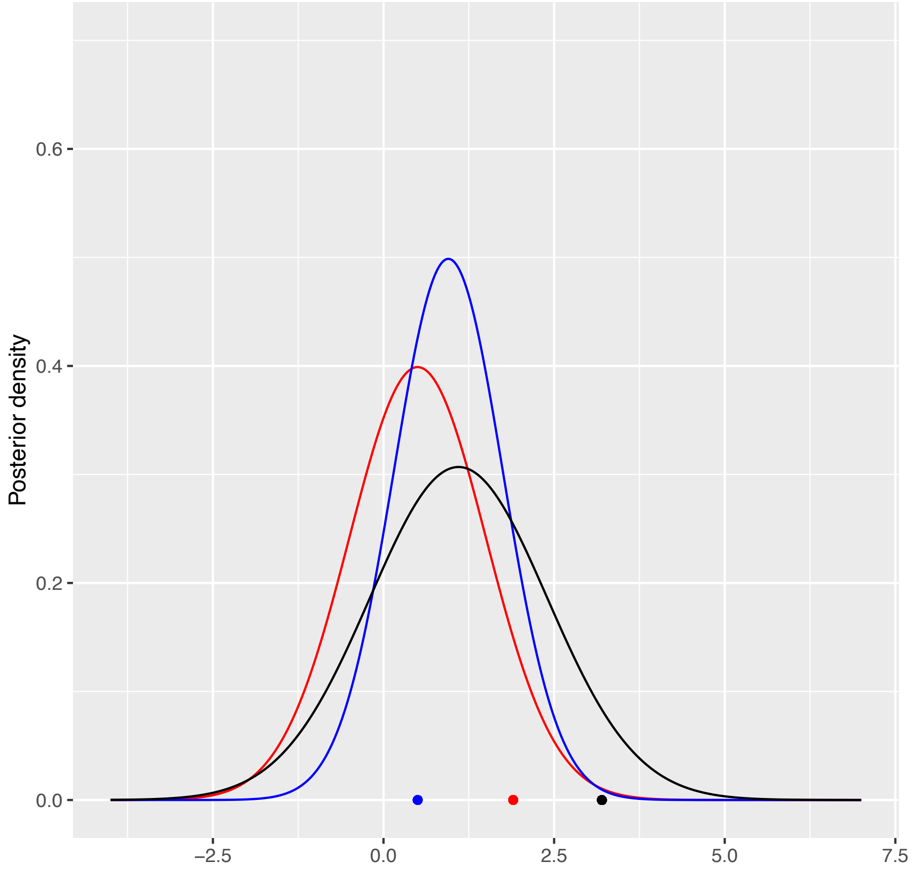
- Consider the posterior distribution on the mean of each action.
- Draw a sample from each and deploy action with the highest sample.
- Observe the actual X_k and update posterior.



More Successful Strategies

Randomised Approach

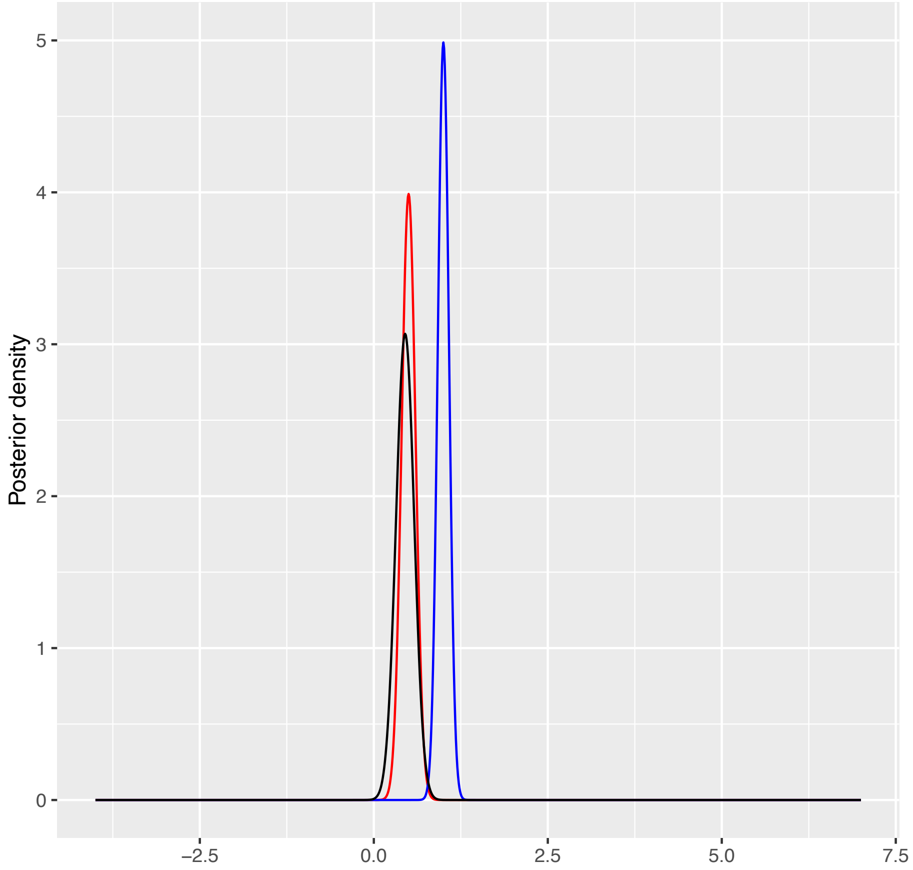
- Consider the posterior distribution on the mean of each action.
- Draw a sample from each and deploy action with the highest sample.
- Observe the actual X_k and update posterior.
- Repeat.



More Successful Strategies

Randomised Approach

- Consider the posterior distribution on the mean of each action.
- Draw a sample from each and deploy action with the highest sample.
- Observe the actual X_k and update posterior.
- Repeat.
- Eventually distributions concentrate.



General Framework

- These methods (appropriately tuned) are successful for the multi-armed bandit problem - and for many more general online problems.
- Performance is measured by **regret**.
- Suppose in each round we choose $A_t \in \mathcal{A} \quad (= \{1, \dots, K\}$ for MAB)
- And then observe reward $R(A_t) \quad (= X_{A_t}$ for MAB)
- Let the optimal action be $A^* = \text{argmax}_{A \in \mathcal{A}} E(R(A_t))$

$$Reg(T) = \sum_{t=1}^{T} E(R(A^*) - R(A_t)) = T \cdot E(R(A^*)) - \sum_{t=1}^{T} E(R(A_t))$$

- Theoretical property, analysed in worst case.

Application 1: Perimeter Surveillance

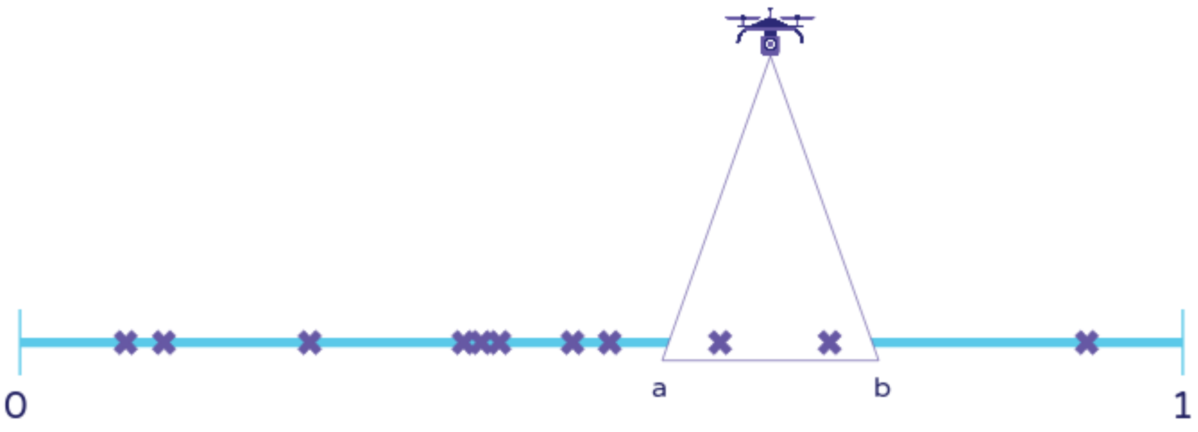
Perimeter Surveillance

- Events happen as a (spatially) inhomogeneous Poisson process
- Sensors can choose which sub-interval (a,b) from the [0, 1] interval to observe

1-dimensional world

Events

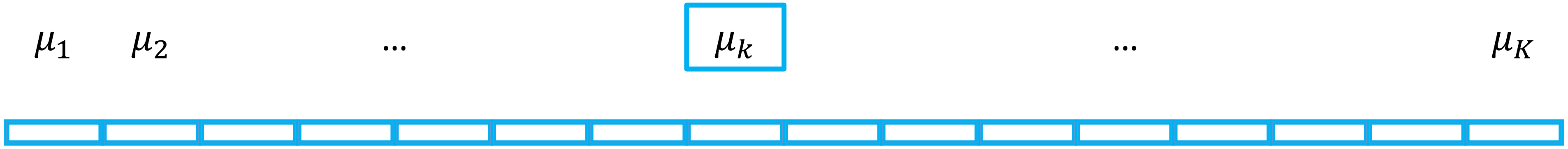
Sensors



Perimeter Surveillance

- We discretise the perimeter, so we have Poisson counts in each cell

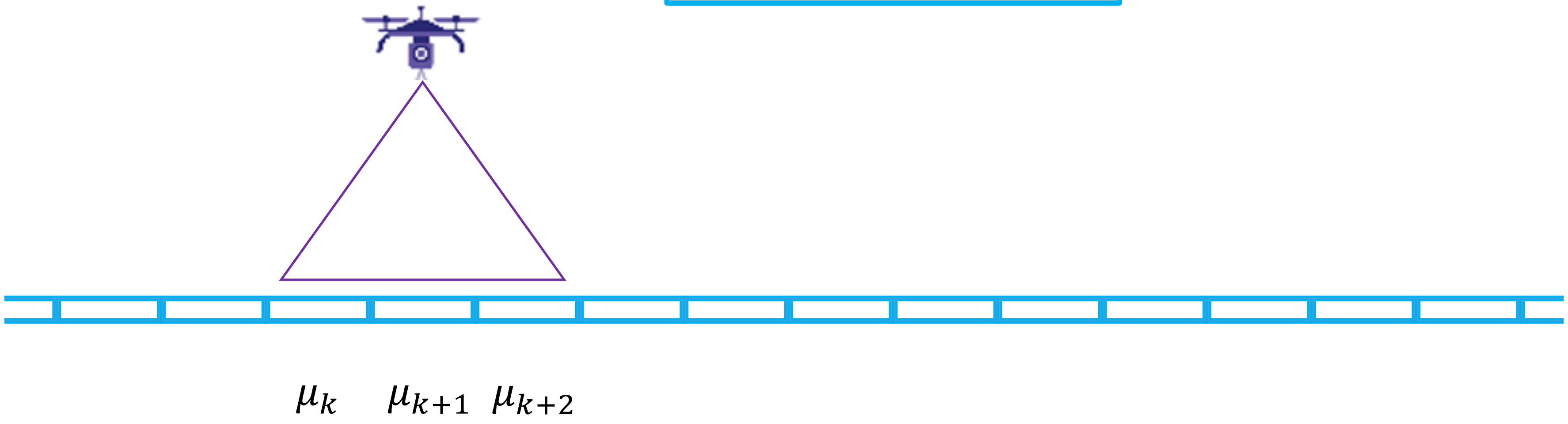
$$X_k \sim Pois(\mu_k)$$



Perimeter Surveillance

- A sensor is deployed to a set of cells, and observes a filtered set of events

$$X_k \sim Pois(\mu_k)$$
$$X_{k+1} \sim Pois(\mu_{k+1})$$
$$X_{k+2} \sim Pois(\mu_{k+2})$$

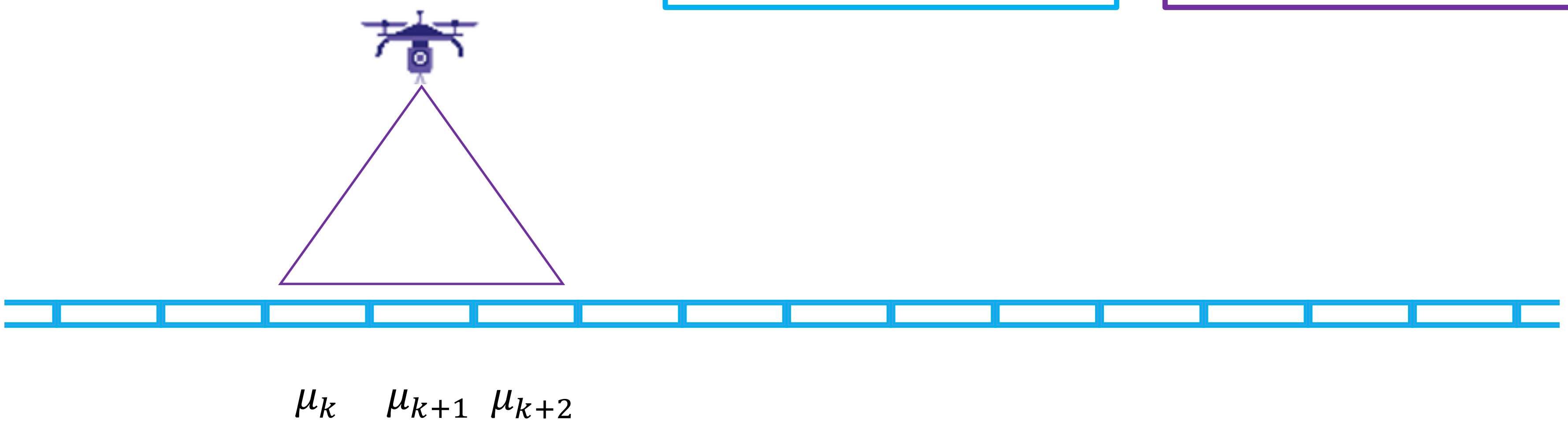


Perimeter Surveillance

- A sensor is deployed to a set of cells, and observes a filtered set of events

$$X_k \sim Pois(\mu_k)$$
$$X_{k+1} \sim Pois(\mu_{k+1})$$
$$X_{k+2} \sim Pois(\mu_{k+2})$$

Choice of set $\{k, \ldots, k+2\}$, gives rise to filtering probability $\gamma(3) \in [0,1]$



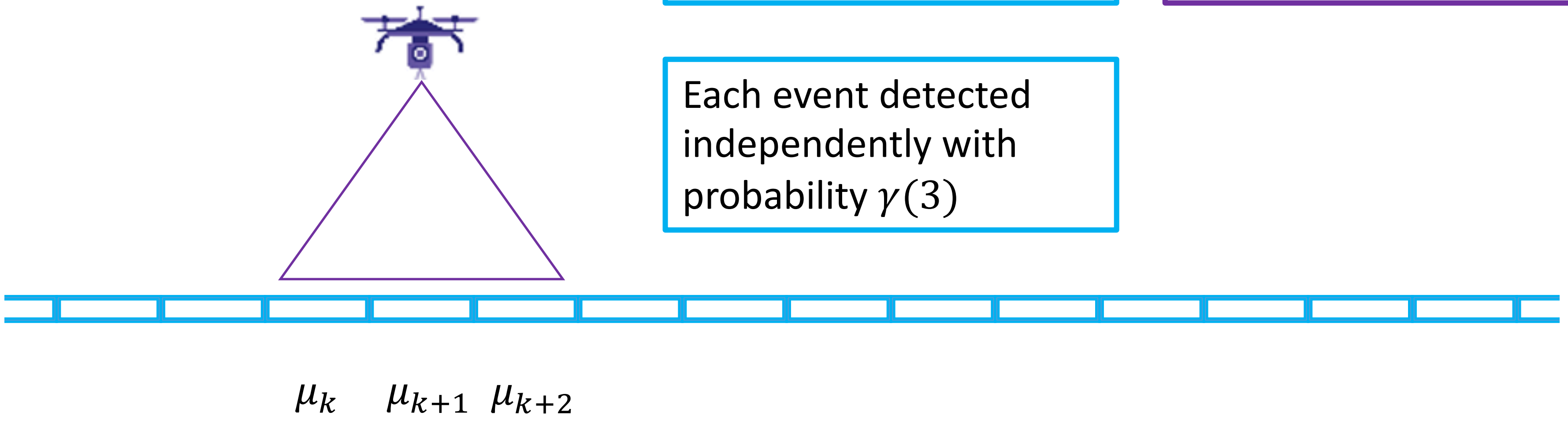
Perimeter Surveillance

- A sensor is deployed to a set of cells, and observes a filtered set of events

$$X_k \sim Pois(\mu_k)$$
$$X_{k+1} \sim Pois(\mu_{k+1})$$
$$X_{k+2} \sim Pois(\mu_{k+2})$$

Choice of set $\{k, \ldots, k+2\}$, gives rise to filtering probability $\gamma(3) \in [0,1]$

Each event detected independently with probability $\gamma(3)$



Perimeter Surveillance

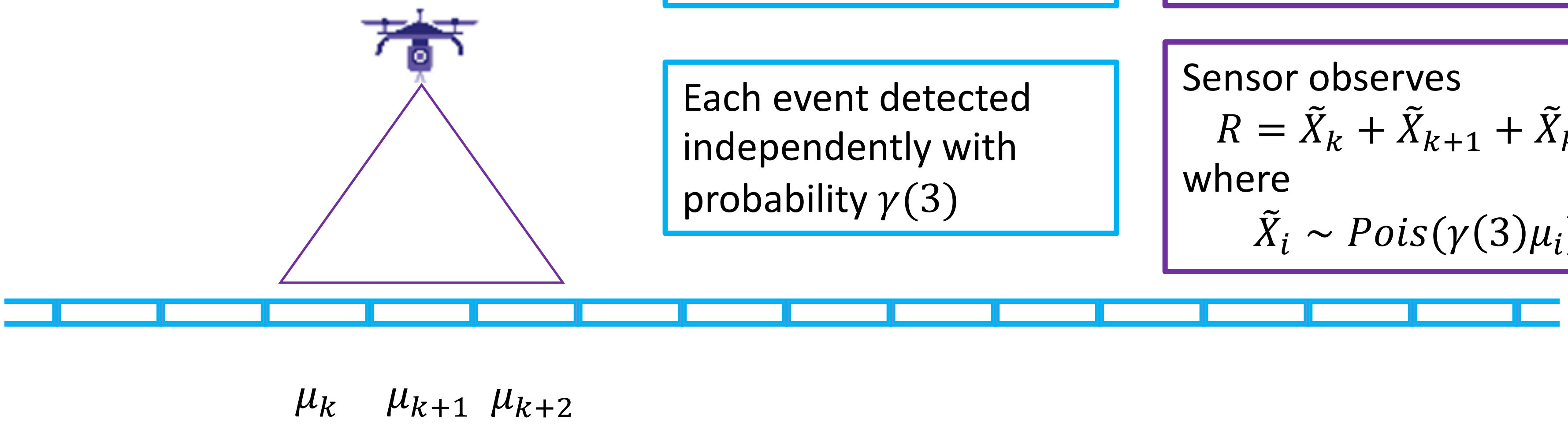
- A sensor is deployed to a set of cells, and observes a filtered set of events

$$X_k \sim Pois(\mu_k)$$
$$X_{k+1} \sim Pois(\mu_{k+1})$$
$$X_{k+2} \sim Pois(\mu_{k+2})$$

Choice of set $\{k, \dots, k+2\}$, gives rise to filtering probability $\gamma(3) \in [0,1]$

Each event detected independently with probability $\gamma(3)$

Sensor observes
$$R = \tilde{X}_k + \tilde{X}_{k+1} + \tilde{X}_{k+2}$$
where
$$\tilde{X}_i \sim Pois(\gamma(3)\mu_i)$$



Model

- Let there be K cells and $U < K$ sensors.
- Let the sensor u have filtering probability function γ_u
- Let $\boldsymbol{a}_u \subset \{1, \dots K\}$ be the cells assigned to sensor u
- We wish to (learn to) optimise

$$\max_{\boldsymbol{a}_u, u=1,\dots,U} \sum_{u=1}^{U} \gamma_u(|\boldsymbol{a}_u|) \sum_{k \in \boldsymbol{a}_u} \mu_k$$

$$s.t.\ \ \boldsymbol{a}_u \cap \boldsymbol{a}_v = \emptyset, \forall\, u \neq v$$

Solution Approaches

- We compute the solution to an optimisation of the form below at each $t \in \{1, \ldots, T\}$

$$\boldsymbol{a}_t = \underset{\boldsymbol{a}_u, u=1,\ldots,U}{\operatorname{argmax}} \sum_{u=1}^{U} \gamma_u(|\boldsymbol{a}_u|) \sum_{k \in \boldsymbol{a}_u} \mu_k$$

$$s.t. \;\; \boldsymbol{a}_u \cap \boldsymbol{a}_v = \emptyset, \forall\, u \neq v$$

Solution Approaches

- We compute the solution to an optimisation of the form below at each $t \in \{1, \dots, T\}$

$$\boldsymbol{a}_t = \underset{\boldsymbol{a}_u, u=1,\dots,U}{\operatorname{argmax}} \sum_{u=1}^{U} \gamma_u(|\boldsymbol{a}_u|) \sum_{k \in \boldsymbol{a}_u} \boldsymbol{\mu_k}$$

$$s.t.\ \ \boldsymbol{a}_u \cap \boldsymbol{a}_v = \emptyset, \forall\, u \neq v$$

- Since μ_k are unknown we replace them with optimistic or randomised estimates

Estimate Design

- The observed data for a bin k is a series of Poisson r.v.s with means $\gamma_1\mu_k, \gamma_2\mu_k, \ldots, \gamma_N\mu_k$ for some sequence $\gamma_1, \gamma_2, \ldots, \gamma_N \in [0,1]^N$.
- For a randomised approach, Gamma prior is conjugate, so posterior sampling is straightforward.
  - Replace μ_k with a sample from a Gamma posterior.
- For an optimistic approach, non-independence raises complexities.
  - Can use martingale inequalities to derive upper confidence bound:

$$P\left(\hat{\mu}_{k,N} + C_N \geq \mu_k\right) \geq 1 - \delta$$

Optimistic Strategy (UCB)

- Initial phase: choose actions randomly to initialize mean estimates.
- Iterative phase, at each time $t \leq T$
  - Compute mean estimate for each bin $\hat{\mu}_{k,t}$
  - Compute the upper confidence bound term

$$C_{k,t} = O\left(\sqrt{\frac{\mu_{k,max}}{\Gamma_{k,t}}}\right)$$

  - Choose an action which is optimal w.r.t the upper confidence bounds,

$$\boldsymbol{a}_t = \underset{\boldsymbol{a}_u, u=1,\dots,U}{\operatorname{argmax}} \sum_{u=1}^{U} \gamma_u(|\boldsymbol{a}_u|) \sum_{k \in \boldsymbol{a}_u} (\hat{\mu}_{k,t} + C_{k,t})$$

$$s.t. \;\; \boldsymbol{a}_u \cap \boldsymbol{a}_v = \emptyset, \forall\, u \neq v$$

$\mu_{k,max}$: upper bound on μ_k

$\Gamma_{k,t}$: sum of detection probability in k so far

Randomised Strategy (Thompson Sampling)

- Initialise via a Gamma prior on each mean parameter $\mu_k \sim Gamma(\alpha_k, \beta_k)$.
- Iterative phase, at each time $t \leq T$
  - Draw a sample from the posterior for each bin,

$$\tilde{\mu}_{k,t} \sim Gamma(\alpha_k + S_{k,t}, \beta_k + \Gamma_{k,t})$$

  - Choose an action which is optimal w.r.t the Thompson Samples,

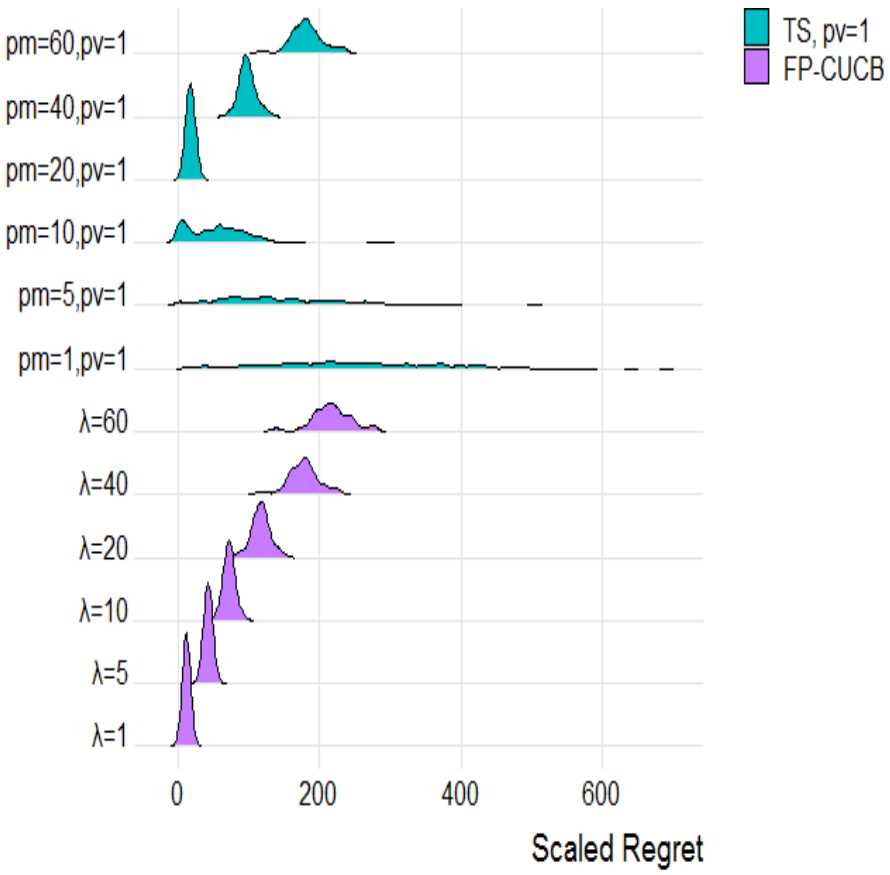
$$\boldsymbol{a}_t = \underset{\boldsymbol{a}_u, u=1,\dots,U}{\operatorname{argmax}} \sum_{u=1}^{U} \gamma_u(|\boldsymbol{a}_u|) \sum_{k \in \boldsymbol{a}_u} \tilde{\mu}_{k,t}$$

$$s.t.\ \ \boldsymbol{a}_u \cap \boldsymbol{a}_v = \emptyset, \forall\, u \neq v$$

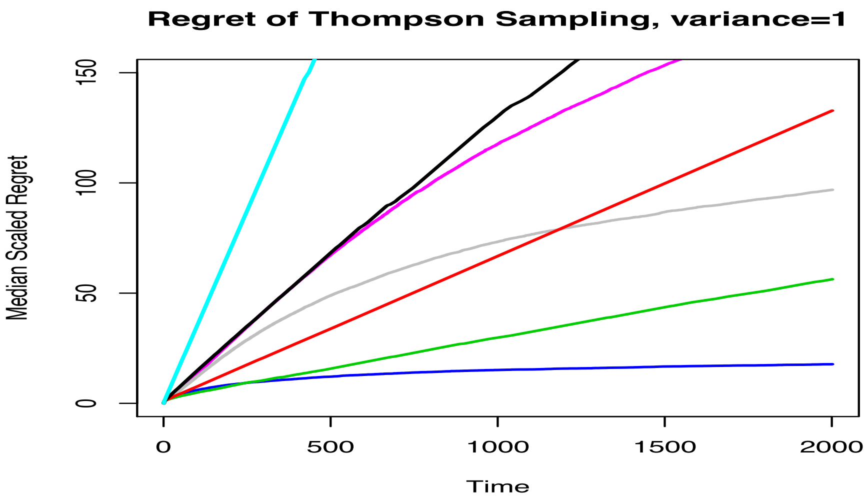
$S_{k,t}$: sum of events observed in k so far

$\Gamma_{k,t}$: sum of detection probability in k so far

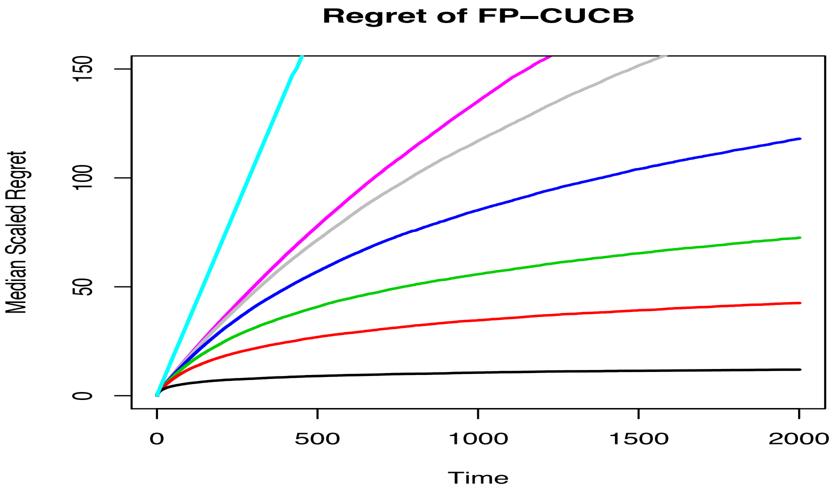
Results



pm=60,pv=1
pm=40,pv=1
pm=20,pv=1
pm=10,pv=1
pm=5,pv=1
pm=1,pv=1
λ=60
λ=40
λ=20
λ=10
λ=5
λ=1
0
200
400
600
Scaled Regret
TS, pv=1
FP-CUCB



Regret of Thompson Sampling, variance=1
Median Scaled Regret
0
50
100
150
0
500
1000
1500
2000
Time



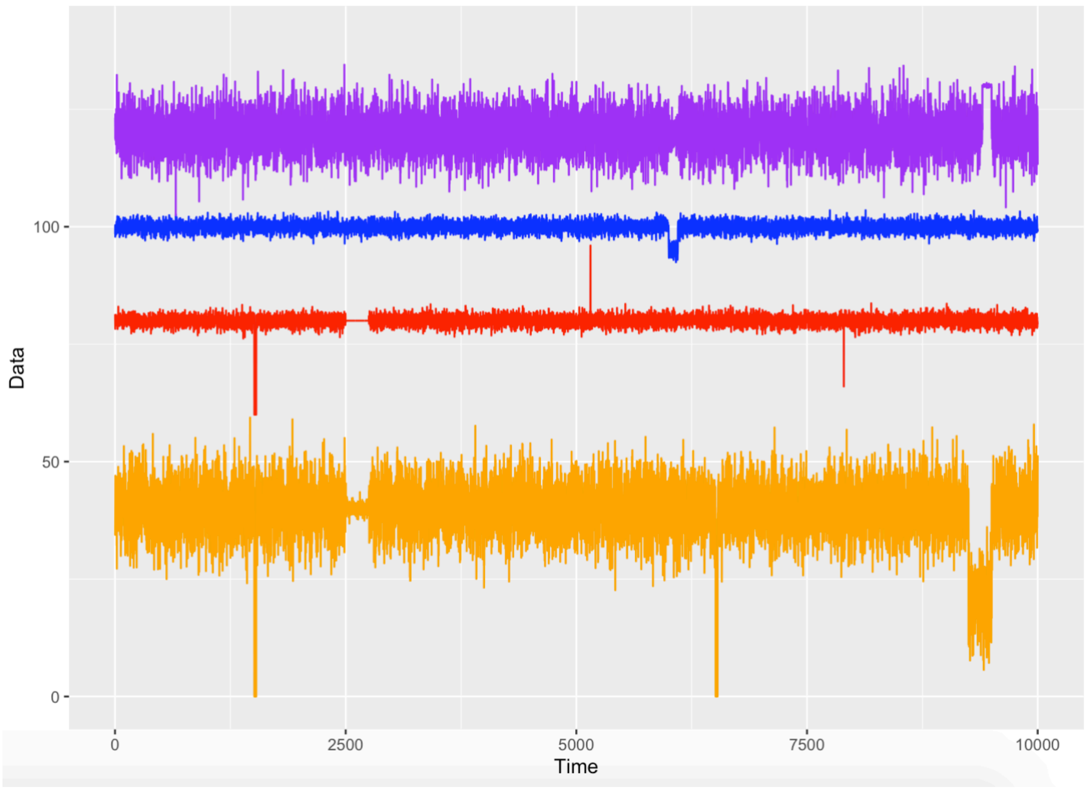
Regret of FP–CUCB
Median Scaled Regret
0
50
100
150
0
500
1000
1500
2000
Time

Application 2: Quality Control

Apple Tasting Model

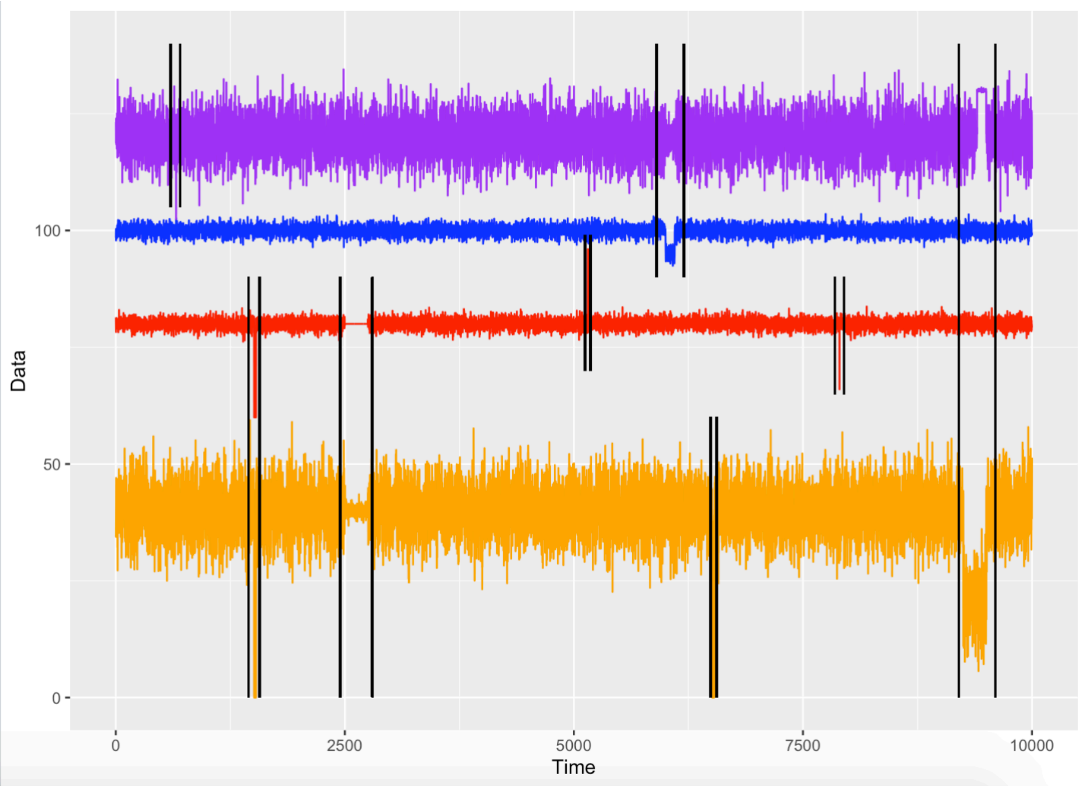
- A new apple presented at time t
- It is either
  - Class 0: good
  - Class 1: rotten
- We want to remove rotten apples, and let good apples pass
- We can only tell the class by removing and tasting.
- Need to balance tasting/passing

Network Traffic Data



- Monitoring a set of data streams
- Occasional anomalies occur, which are either
  - Class 0 – innocuous
  - Class 1 – relevant

Network Traffic Data



- Monitoring a set of data streams
- Occasional anomalies occur, which are either
  - Class 0 – innocuous
  - Class 1 – relevant
- A time series algorithm flags these, and we can determine the class by showing to an engineer
- Showing an engineer entails a cost, and therefore we only want to display relevant anomalies

Quality Control

- Points where the time series algorithm flags an anomaly become decision times.
- We want to learn the parameters of a classifier.
- Assume a logistic regression model

$$P(C_t = 1) = \sigma(x_t^\top \theta^*) = \frac{\exp(-x_t^\top \theta^*)}{1 + \exp(-x_t^\top \theta^*)}$$

- x_t are features of the anomaly
- θ^* is an unknown parameter vector

Randomised Approach (Thompson Sampling)

- Potential anomaly proposed with feature vector x_t
- Draw a sample $\tilde{\theta}_t$ from the posterior* π_t on parameter θ^*
- Estimate the probability of being a relevant anomaly

$$\tilde{p}_t = \sigma\left(x_t^\top \tilde{\theta}_t\right).$$

- Display to engineer if expected cost is minimised by doing so.
- If displayed to engineer, receive true class as feedback.
- In either case, incur cost (unknown (to algorithm) if not displayed).

Randomised Approach (Thompson Sampling)

- Potential anomaly proposed with feature vector x_t
- Draw a sample $\tilde{\theta}_t$ from the posterior* π_t on parameter θ^*
- Estimate the probability of being a relevant anomaly

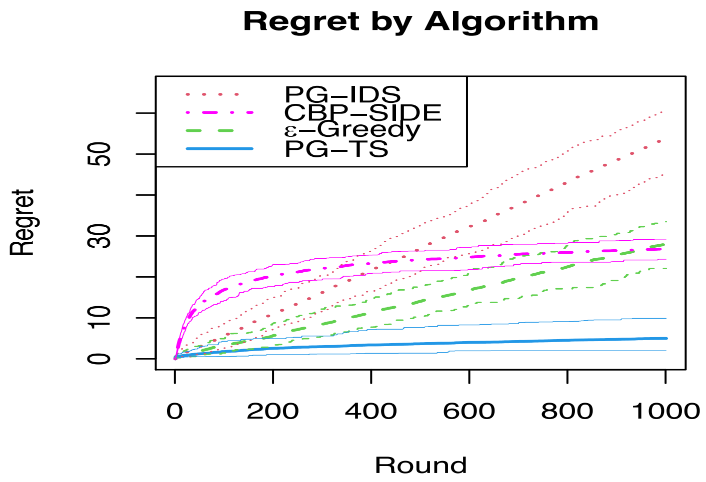
$$\tilde{p}_t = \sigma\left(x_t^\top \tilde{\theta}_t\right).$$

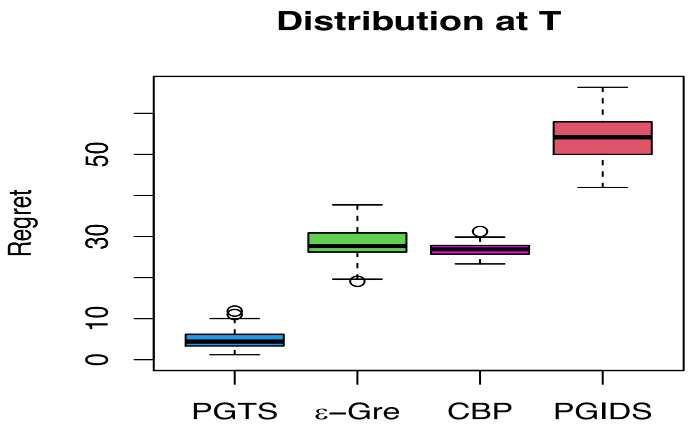
- Display to engineer if expected cost is minimised by doing so.
- If displayed to engineer, receive true class as feedback.
- In either case, incur cost (unknown (to algorithm) if not displayed).

*We require an approximation to the posterior – but the approximation is consistent in a limiting sense.

Performance of Thompson Sampling

- Compare against:
  - IDS (a hybrid of optimisation and randomisation)
  - CBP-side (an optimistic approach)
  - ϵ –Greedy (exploration is independent of the data)





Summary

- Online learning, benefits of optimism and randomisation
- Applications in surveillance, and quality control
- Papers (will) contain theoretical analysis of regret - showing the optimality of these approaches.

Thank you

-

j.grant@lancaster.ac.uk

- Grant, J.A., Leslie, D.S., Glazebrook, K., Szechtman, R., and Letchford, A.N. (2019). Adaptive Policies for Perimeter Surveillance Problems. *European Journal of Operational Research.*
- Grant, J.A., Leslie D.S. (2020). Apple Tasting Revisited: Partially Monitored Online Binary Classification. *Working Paper.*