Trying to make the right decision when you can't see everything

Yoni Nazarathy

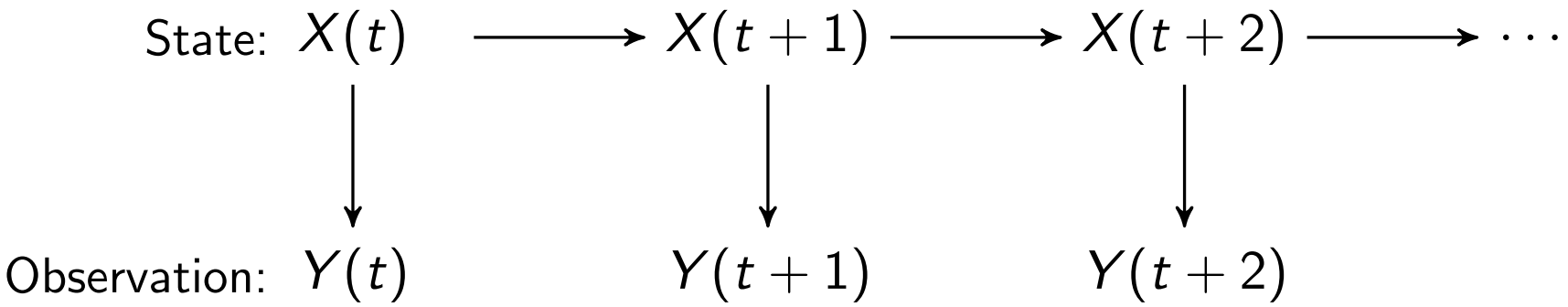
The University of Queensland

Sep 21, 2015

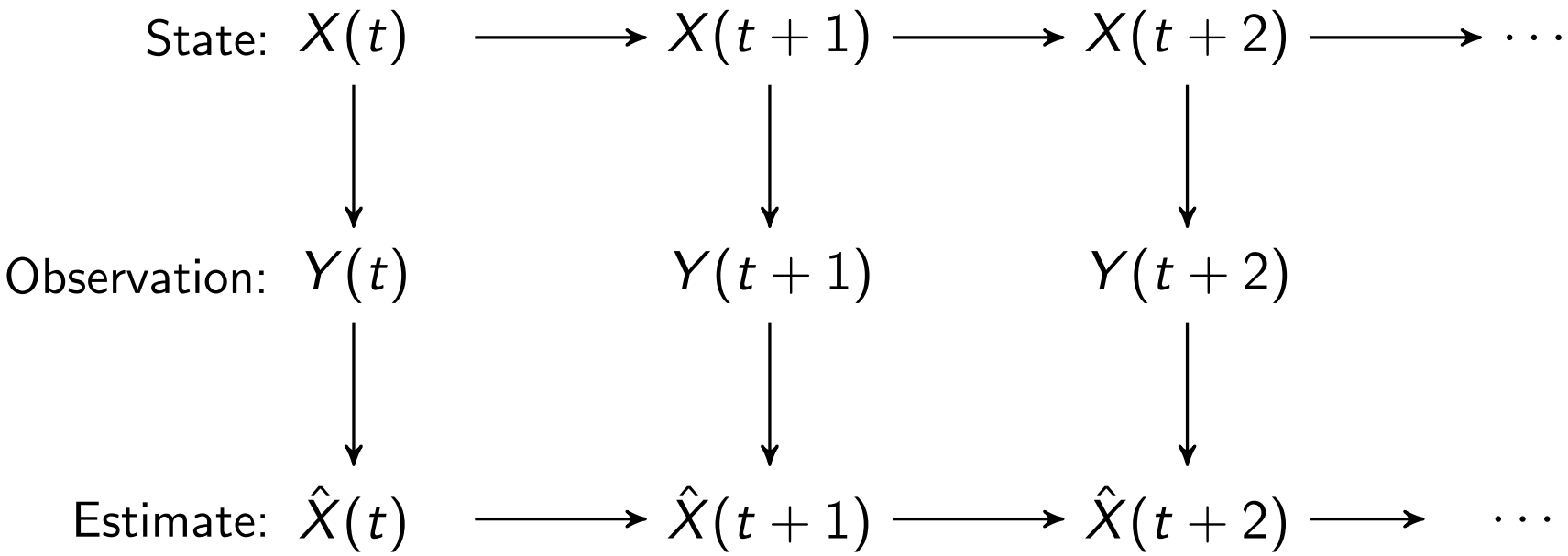
Discrete Time Setup

State: $X(t) \longrightarrow X(t+1) \longrightarrow X(t+2) \longrightarrow \cdots$

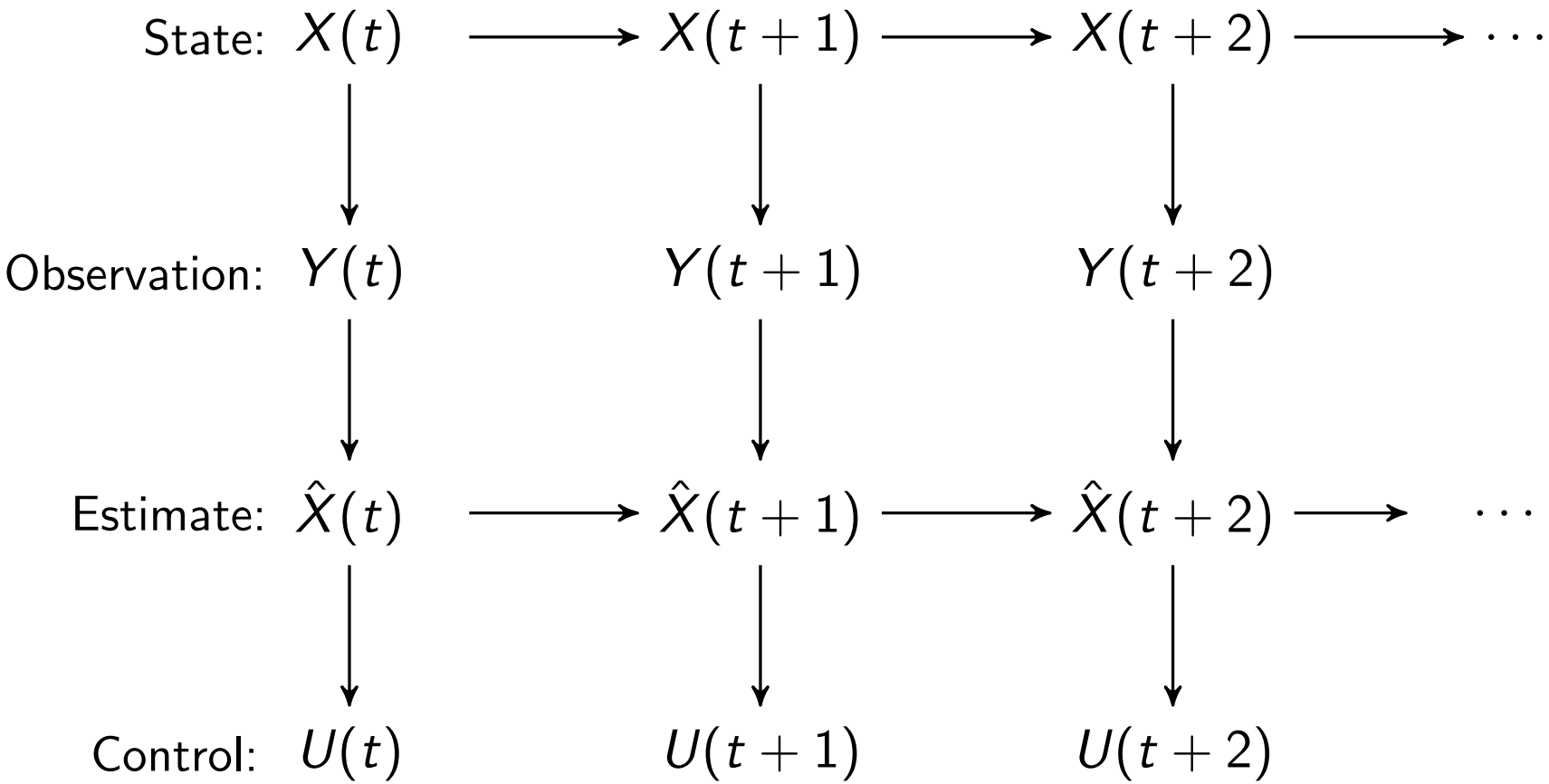
Discrete Time Setup



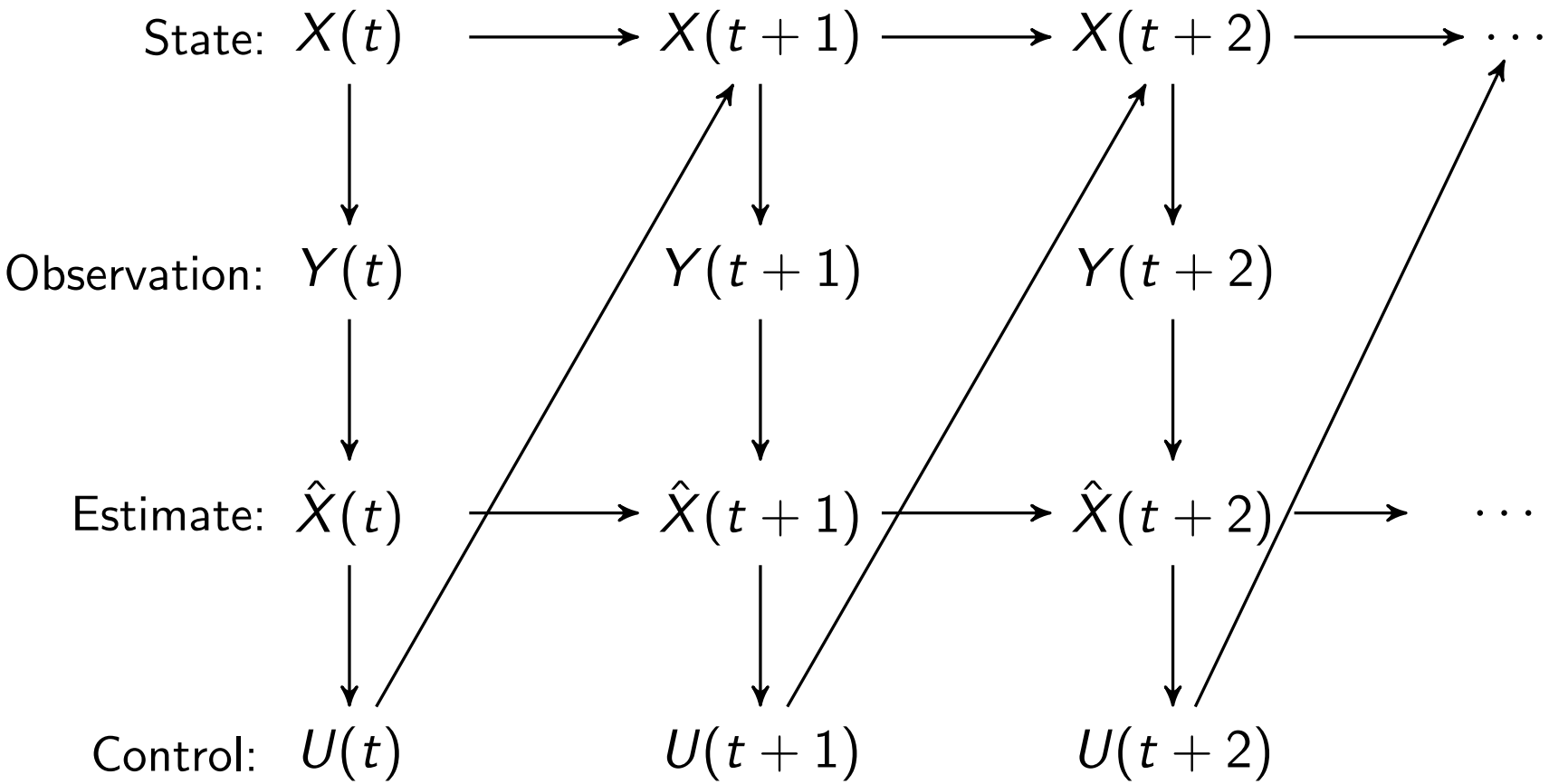
Discrete Time Setup



Discrete Time Setup



Discrete Time Setup

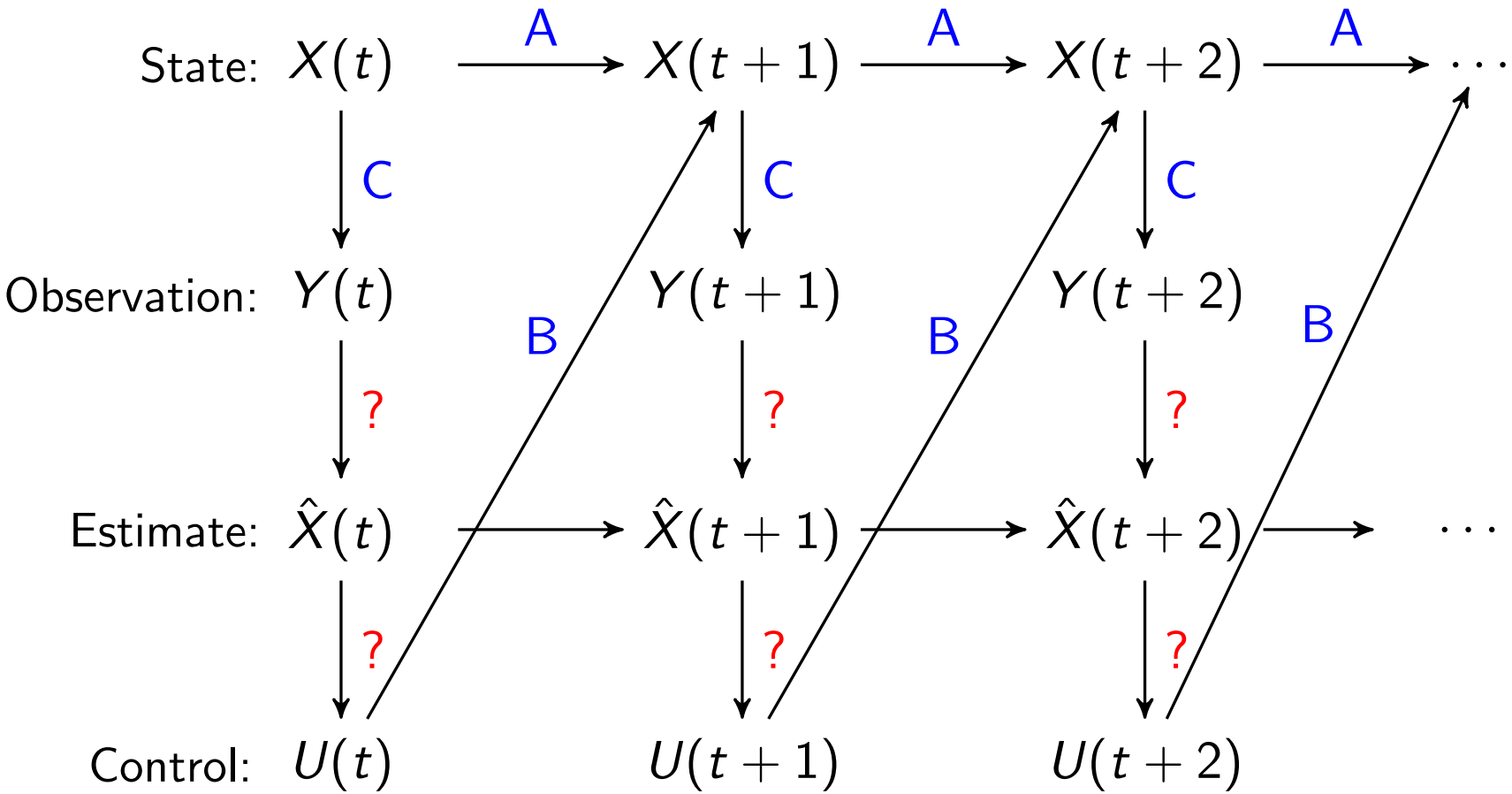


Outline

- Linear Systems: A Success Story
- Stabilising Control of Queues with Hidden Environments
- Reward Observing Restless Bandits

(Deterministic) Linear Systems

(Deterministic) Linear Systems Setup



$$X(t+1) = A\,X(t) + B\,U(t) \qquad Y(t) = C\,X(t)$$

The Luenberger Observer

$$\hat{X}(t) = A\hat{X}(t) + BU(t) - K_o\big(\hat{Y}(t) - Y(t)\big) \quad \text{with} \quad \hat{Y}(t) = C\hat{X}(t)$$

The Luenberger Observer

$$\hat{X}(t) = A\hat{X}(t) + BU(t) - K_o\big(\hat{Y}(t) - Y(t)\big) \quad \text{with} \quad \hat{Y}(t) = C\hat{X}(t)$$

$$\begin{aligned} e(t+1) &= X(t+1) - \hat{X}(t+1) \\ &= AX(t) + BU(t) - \Big(A\hat{X}(t) + BU(t) - K_o\big(C\hat{X}(t) - CX(t)\big)\Big) \\ &= AX(t) - A\hat{X}(t) - K_o C\big(X(t) - \hat{X}(t)\big) \\ &= (A - K_o\, C)e(t) \end{aligned}$$

The Luenberger Observer

$$\hat{X}(t) = A\hat{X}(t) + BU(t) - K_o\big(\hat{Y}(t) - Y(t)\big) \quad \text{with} \quad \hat{Y}(t) = C\hat{X}(t)$$

$$\begin{aligned} e(t+1) &= X(t+1) - \hat{X}(t+1) \\ &= AX(t) + BU(t) - \Big(A\hat{X}(t) + BU(t) - K_o\big(C\hat{X}(t) - CX(t)\big)\Big) \\ &= AX(t) - A\hat{X}(t) - K_o C\big(X(t) - \hat{X}(t)\big) \\ &= (A - K_o\, C)e(t) \end{aligned}$$

Would like $e(t) \to 0$

The Luenberger Observer

$$\hat{X}(t) = A\hat{X}(t) + BU(t) - K_o\big(\hat{Y}(t) - Y(t)\big) \quad \text{with} \quad \hat{Y}(t) = C\hat{X}(t)$$

$$\begin{aligned} e(t+1) &= X(t+1) - \hat{X}(t+1) \\ &= AX(t) + BU(t) - \Big(A\hat{X}(t) + BU(t) - K_o\big(C\hat{X}(t) - CX(t)\big)\Big) \\ &= AX(t) - A\hat{X}(t) - K_o C\big(X(t) - \hat{X}(t)\big) \\ &= (A - K_o\, C)e(t) \end{aligned}$$

Would like $e(t) \to 0$

Observer Design: Select K_o so that $\text{sp}(A - K_o\, C) < 1$.
This can always be done if the pair (A, C) satisfies a rank condition (observability).

State Feedback Control and The Separation Principle

As controller would like to take $U(t) = -K_f X(t)$ so that,

$$\begin{aligned} X(t+1) &= AX(t) + BU(t) \\ &= (A - B\,K_f)X(t) \end{aligned}$$

If (A, B) satisfy a rank condition (controllability) then can choose K_f so as to have arbitrary eigenvalues of $(A - B\,K_f)$ (e.g. stabilise).

State Feedback Control and The Separation Principle

As controller would like to take $U(t) = -K_f X(t)$ so that,

$$\begin{aligned} X(t+1) &= AX(t) + BU(t) \\ &= (A - B\,K_f)X(t) \end{aligned}$$

If (A, B) satisfy a rank condition (controllability) then can choose K_f so as to have arbitrary eigenvalues of $(A - B\,K_f)$ (e.g. stabilise).

But we don't have $X(t)$, so instead use $U(t) = -K_f \hat{X}(t)$.

The "Separation Principle": Can design the observer (K_o) and the controller (K_f) separately to achieve desired behaviour.

Making the "Right Decisions"

First choose K_o for a "good" observer. Now the separation principle allows to focus on finding K_f:

Stability

Choose K_f so that $X(t) \to 0$

Making the "Right Decisions"

First choose K_o for a "good" observer. Now the separation principle allows to focus on finding K_f:

Stability

Choose K_f so that $X(t) \to 0$

Quadratic Regulation

$$\min_U \quad \sum_{t=1}^{\infty} X(t)'QX(t) + U(t)'RU(t)$$

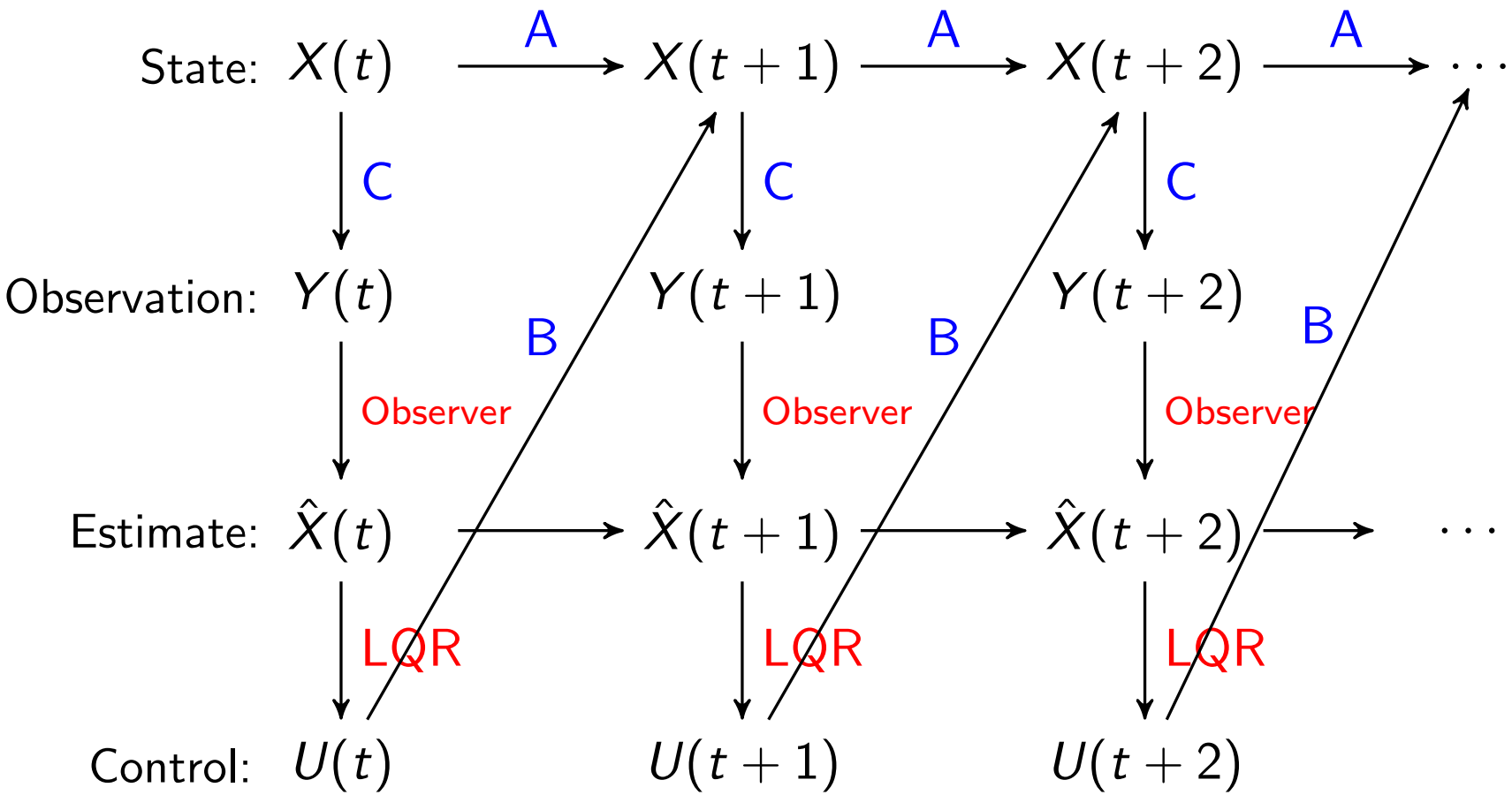
Solution to this problem (LQR):

$$U(t) = -K_f X(t),$$

with K_f based on a solution of a Riccati equation.

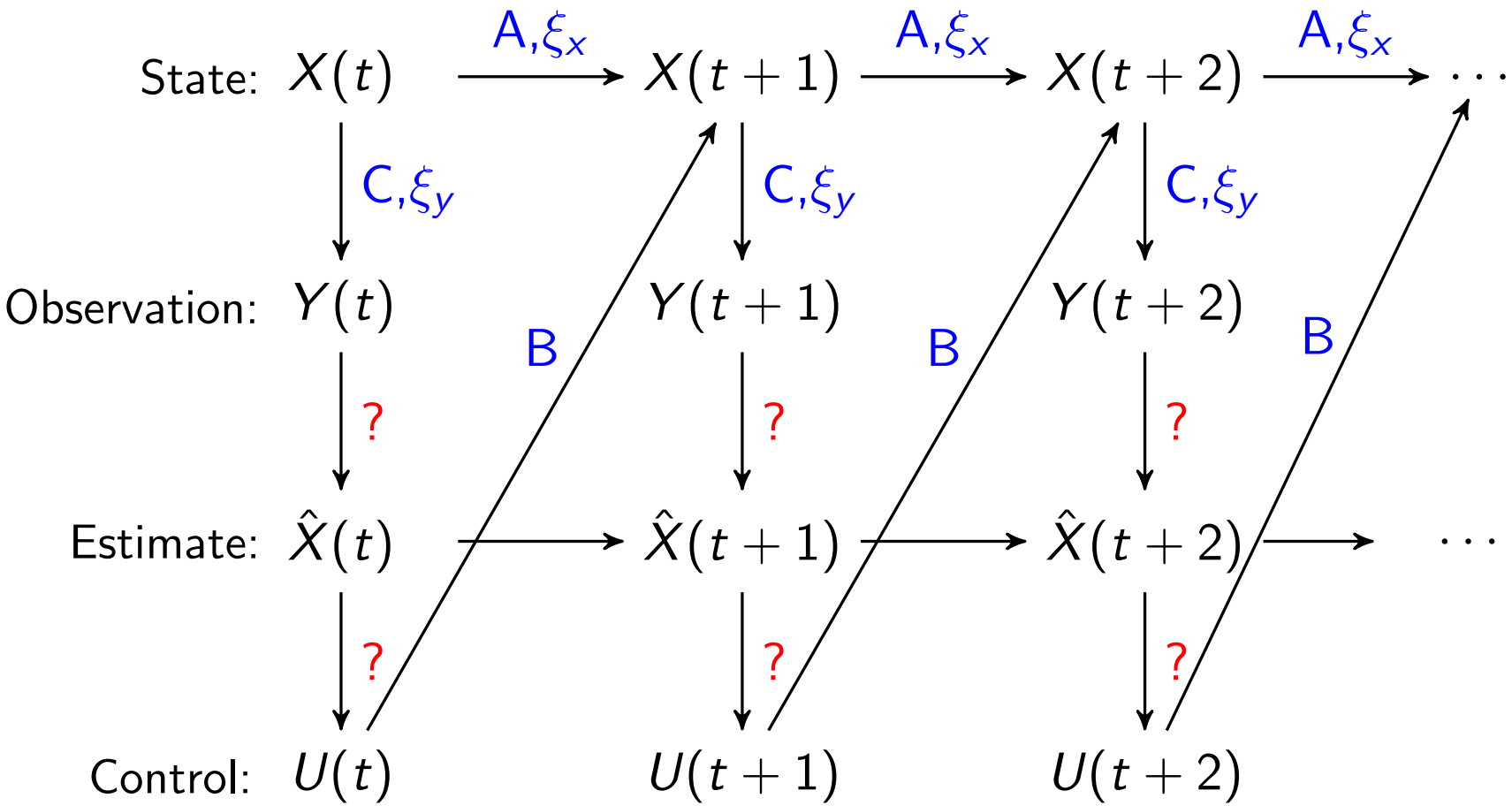
Moral: For such systems: Don't worry about the fact that "you can't see everything" when you choose an optimal decision

(Deterministic) Linear Systems Summary



Linear Systems with Noise (Stochastic)

(Stochastic) Linear Systems Setup



$$X(t+1) = A\,X(t) + B\,U(t) + \xi_x(t) \qquad Y(t) = C\,X(t) + \xi_y(t)$$

The noise components $\xi_x(\cdot)$ and $\xi_y(\cdot)$ are i.i.d. Gaussian

Kalman Filtering

Ignore the control and consider:

$$X(t+1) = AX(t) + \xi_x(t)$$
$$Y(t) = CX(t) + \xi_y(t)$$

Kalman Filtering

Ignore the control and consider:

$$X(t+1) = AX(t) + \xi_x(t)$$
$$Y(t) = CX(t) + \xi_y(t)$$

Given $\underline{Y} = \big(Y(1), , \ldots, Y(T)\big)$ and $X(1)$, it is straightforward to compute the MMSE $\hat{\underline{X}} = h(\underline{X})$ of $\underline{X} = \big(X(1), \ldots, X(T)\big)$:

$$\text{argmin}_h \mathbb{E}\big[||\underline{X} - h(\underline{X})||^2\big] = \mathbb{E}\big[X(1), \ldots, X(T) \mid Y(1), \ldots, Y(T)\big]$$

Kalman Filtering

Ignore the control and consider:

$$X(t+1) = AX(t) + \xi_x(t)$$
$$Y(t) = CX(t) + \xi_y(t)$$

Given $\underline{Y} = \big(Y(1), , \ldots, Y(T)\big)$ and $X(1)$, it is straightforward to compute the MMSE $\hat{\underline{X}} = h(\underline{X})$ of $\underline{X} = \big(X(1), \ldots, X(T)\big)$:

$$\text{argmin}_h \mathbb{E}\big[||\underline{X} - h(\underline{X})||^2\big] = \mathbb{E}\big[X(1), \ldots, X(T) \mid Y(1), \ldots, Y(T)\big]$$

The Kalman Filter is a way to do this recursively (online)

Kalman Filtering (cont)

The (steady state) Kalman filter is a Luenberger observer with parameters optimised for solving the MMSE:

$$\hat{X}(t+1) = A\hat{X}(t) - K_k\big(C\,A\,\hat{X}(t) - Y(t+1)\big).$$

Kalman Filtering (cont)

The (steady state) Kalman filter is a Luenberger observer with parameters optimised for solving the MMSE:

$$\hat{X}(t+1) = A\hat{X}(t) - K_k\big(C\,A\,\hat{X}(t) - Y(t+1)\big).$$

The parameter K_k is calculated as follows:

$$\begin{aligned} S &= \lim_{t\to\infty} Cov\big(X(t+1) - \hat{X}(t+1) \;\big|\; X(t), X(t-1), \ldots, X(1)\big) \\ &= A\Big(S - SC'\big(CSC' + \Sigma_y\big)^{-1}CS\Big)A' + \Sigma_x \end{aligned}$$

Now,

$$K_k = SC'\big(CSC' + \Sigma_y\big)^{-1}$$

Making The Right Decisions (in a Stochastic Setting)

The Separation Principle generalises:

Certainty Equivalence Principle (holds for such systems, but not always)

In making optimal decisions use $\hat{X}(t) = \mathbb{E}[X(t) \mid \text{observations}]$ as though it was $X(t)$.

Making The Right Decisions (in a Stochastic Setting)

The Separation Principle generalises:

Certainty Equivalence Principle (holds for such systems, but not always)

In making optimal decisions use $\hat{X}(t) = \mathbb{E}[X(t) \mid \text{observations}]$ as though it was $X(t)$.

This allows to solve the LQG problem:

Quadratic Regulation with Gaussian Noise (LQG)

$$\min_{U} \quad \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^{T} X(t)'QX(t) + U(t)'RU(t)\Big]$$

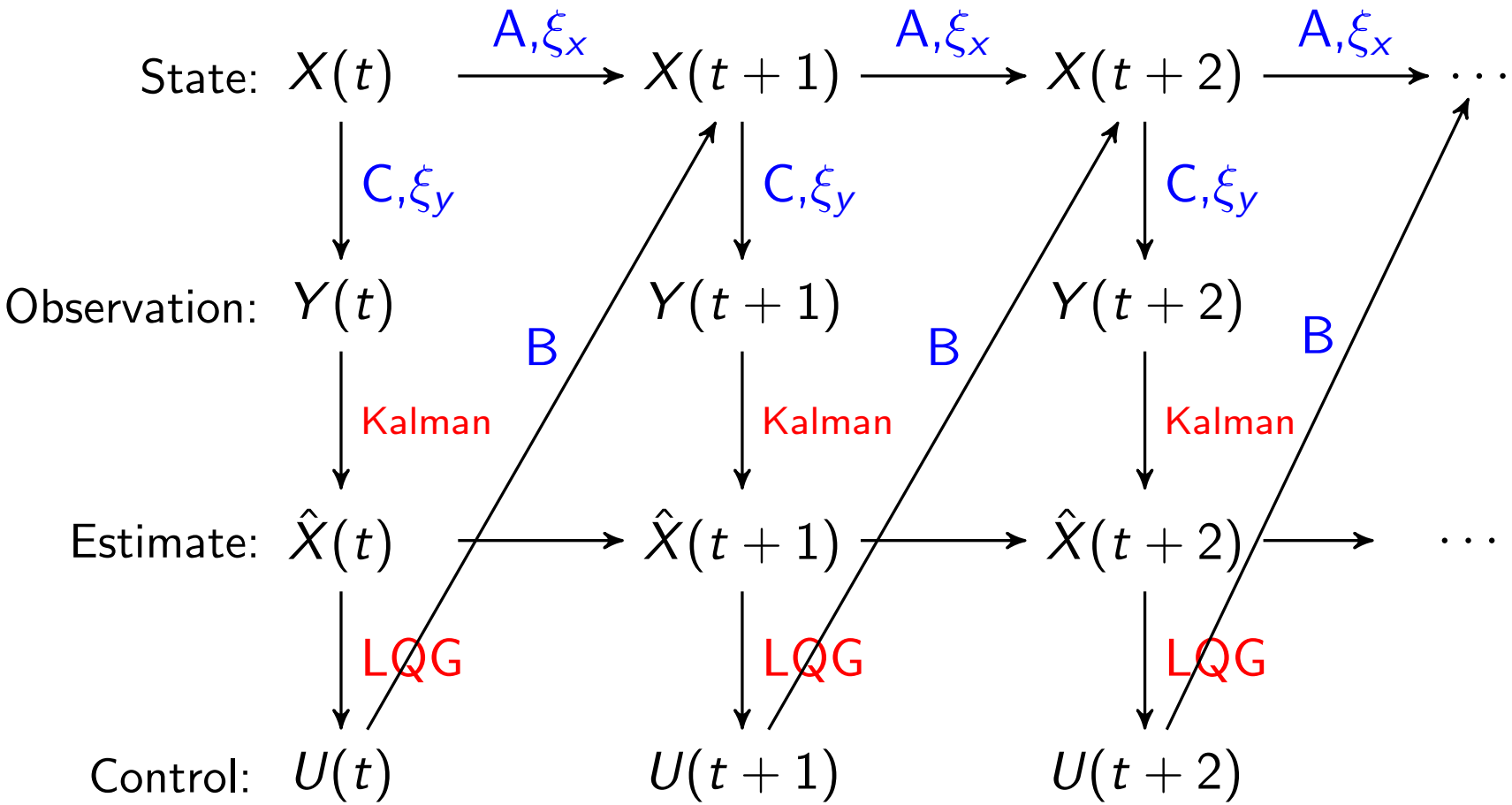
Solution:

$$U(t) = -K_f X(t),$$

with K_f based on a solution of a Riccati equation.

Moral: For such systems don't worry about the fact that "you can't see everything" when you choose an optimal decision.

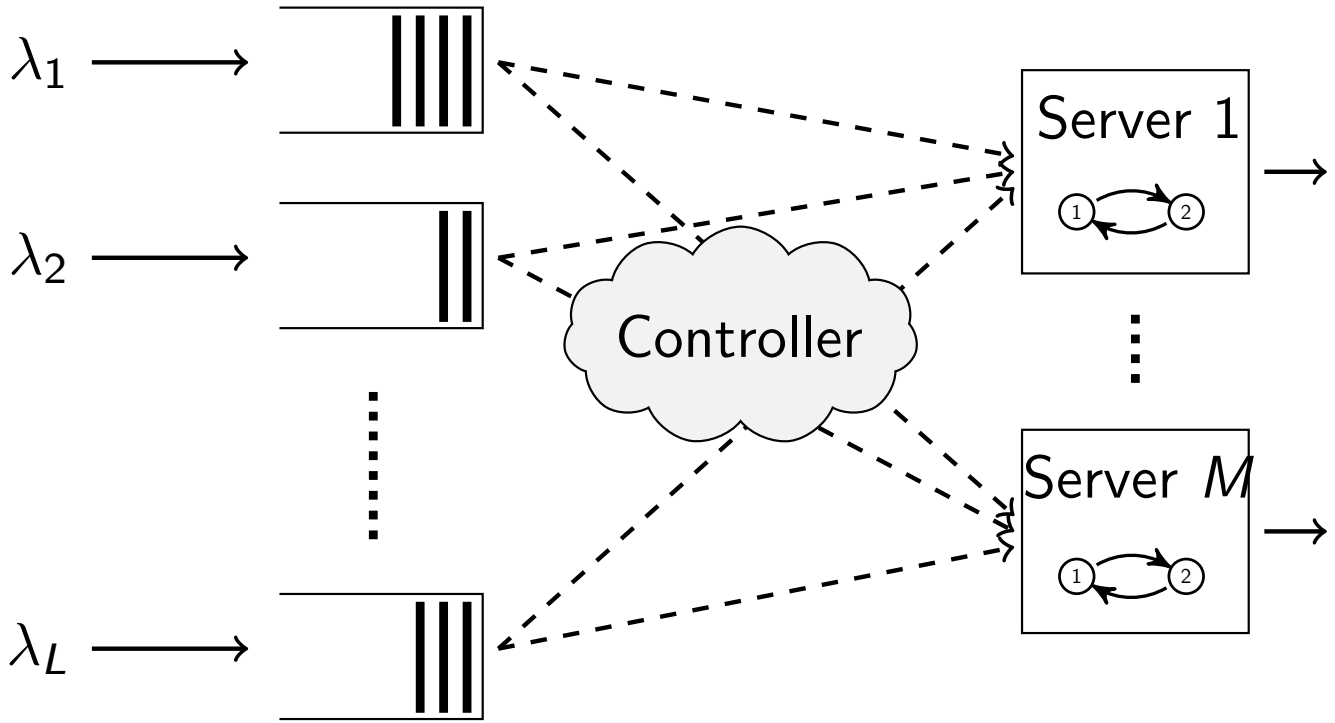
(Stochastic) Linear Systems Summary



Stabilising Control of Queues with Hidden Server States

(preliminary results from a conference paper with T. Taimre, A. Asanjarani, J. Kuhn, B. Patch and A. Vuorinen)

Who Should Serve What?

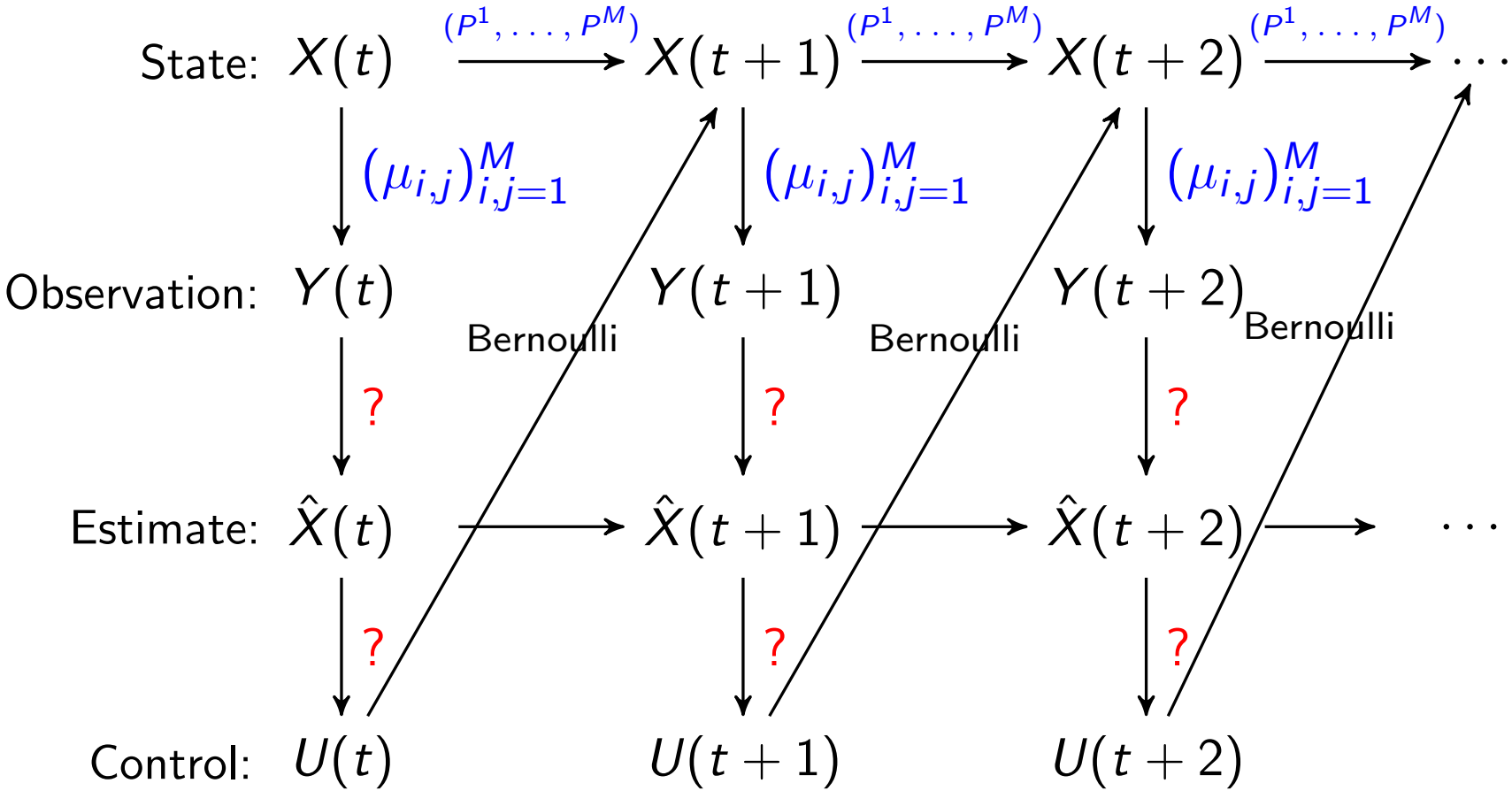


Hidden server states: $X(t) = (x_1(t), \ldots, x_M(t))$ with $x_j(t)$ following a 2 state MC:

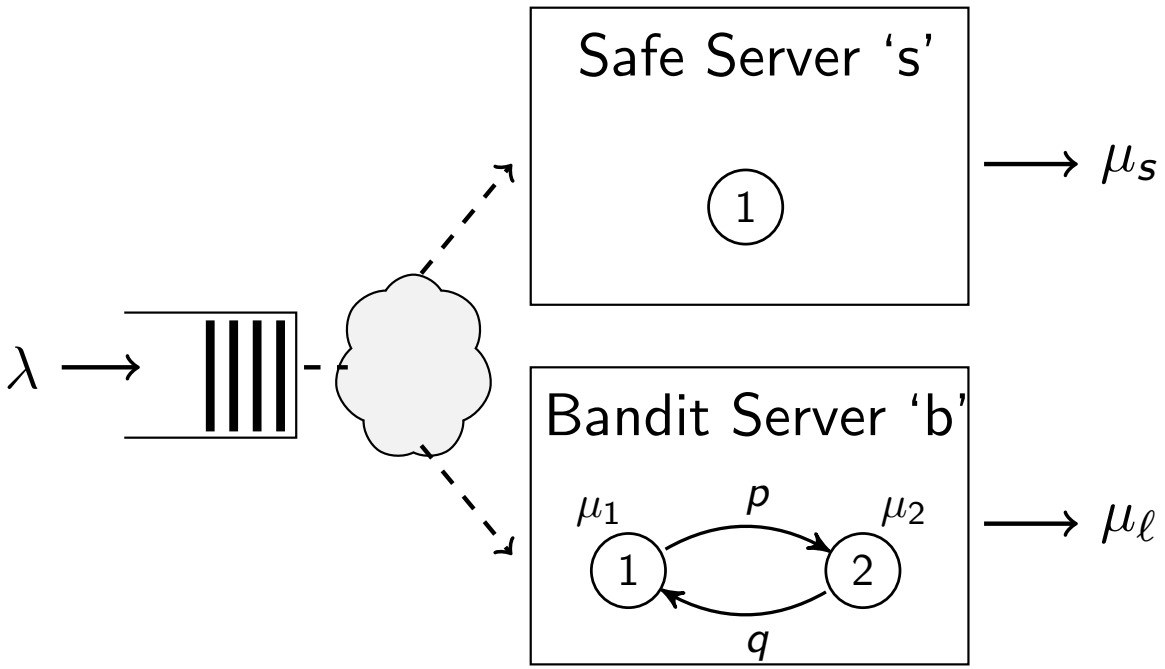
$$P^j = \begin{bmatrix} \bar{p} & p \\ q & \bar{q} \end{bmatrix} = \begin{bmatrix} 1-\gamma\,\bar{\rho} & \gamma\,\bar{\rho} \\ \bar{\gamma}\,\bar{\rho} & 1-\bar{\gamma}\,\bar{\rho} \end{bmatrix}.$$

Service rates are $\mu_{i,j}\big(x_j(t)\big)$. The observations are service successes/failures as well as queue lengths.

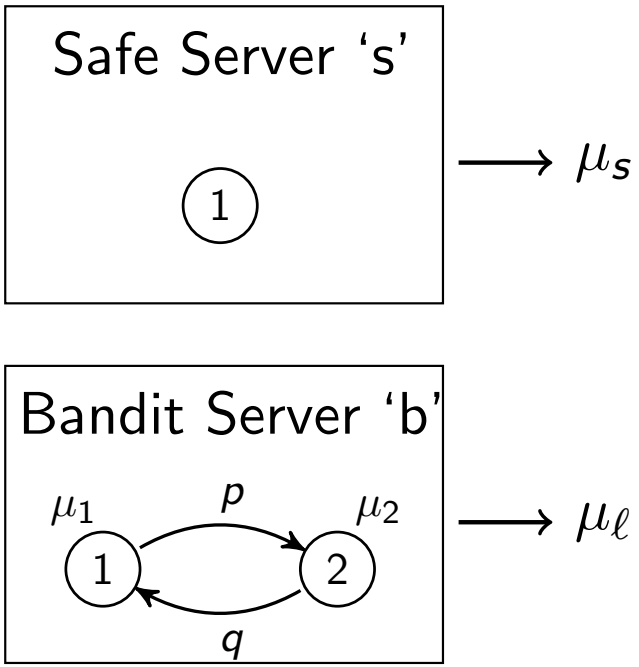
The Setup


State:
$X(t)$
$(P^1, \ldots, P^M)$
$X(t+1)$
$(P^1, \ldots, P^M)$
$X(t+2)$
$(P^1, \ldots, P^M)$
$\cdots$
$(\mu_{i,j})_{i,j=1}^M$
$(\mu_{i,j})_{i,j=1}^M$
$(\mu_{i,j})_{i,j=1}^M$
Observation:
$Y(t)$
$Y(t+1)$
$Y(t+2)$
Bernoulli
Bernoulli
Bernoulli
?
?
?
Estimate:
$\hat{X}(t)$
$\hat{X}(t+1)$
$\hat{X}(t+2)$
$\cdots$
?
?
?
Control:
$U(t)$
$U(t+1)$
$U(t+2)$

A Simpler Problem: Who Should Serve the Queue?



A Similar (Simplest) Problem: Go "Safe" or "Bandit"?



- Independent i.i.d. Bernoulli sequences, $\tilde{Y}_1(t)$, $\tilde{Y}_s(t)$, $\tilde{Y}_2(t)$ with means μ_1, μ_2 and μ_s, respectively, and $\mu_1 < \mu_s < \mu_2$
- $X(t)$ is a 2 state Markov chain
- Observations: $Y(t) = \mathbb{1}\{U(t) = \text{'s'}\}\tilde{Y}_s(t) + \mathbb{1}\{U(t) = \text{'b'}\}\tilde{Y}_{X(t)}$
- Causal policy $U(t)$ should maximize: $\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\big[\sum_{t=0}^{T} Y(t)\big]$

Belief States (POMDP)

Instead of state estimate, $\hat{X}(t)$, keep the conditional distribution (or parameters thereof) of $X(t)$ given history:

$$\omega(t) = \mathbb{P}(X(t) = 2 \mid \text{previous observations and actions})$$

Belief States (POMDP)

Instead of state estimate, $\hat{X}(t)$, keep the conditional distribution (or parameters thereof) of $X(t)$ given history:

$$\omega(t) = \mathbb{P}(X(t) = 2 \mid \text{previous observations and actions})$$

Now based on "no observation", "observation of failure" or "observation of success", update $\omega(t)$ using one of:

$$\tau(\omega) = \omega\rho + \gamma(1-\rho) \qquad \text{(no observation on bandit)}$$

$$\tau_0(\omega) = \frac{\bar{q}\bar{\mu}_2\omega + p\bar{\mu}_1\bar{\omega}}{\bar{\mu}_2\omega + \bar{\mu}_1\bar{\omega}} \qquad \text{(observation of failure on bandit)}$$

$$\tau_1(\omega) = \frac{\bar{q}\mu_2\omega + p\mu_1\bar{\omega}}{\mu_2\omega + \mu_1\bar{\omega}} \qquad \text{(observation of success on bandit)}$$

Value Function and Optimal Decision

For simplicity consider the discounted case with factor $\beta \in (0,1)$:

$$\max \quad \mathbb{E}[\sum_{t=0}^{\infty} \beta^t Y(t)]$$

Bellman's Optimality Equation

$$V(\omega) = \max \Bigg\{ \underbrace{\overbrace{\mu}^{\text{reward}} + \overbrace{\beta V(\tau(\omega))}^{\text{next step}}}_{\text{safe}} \quad , \quad \underbrace{\underbrace{\overbrace{\omega\mu_2}^{\text{good}} + \overbrace{\bar{\omega}\mu_1}^{\text{bad}}}_{\text{reward}} + \beta \overbrace{\left(\underbrace{(\bar{\omega}\bar{\mu}_1 + \omega\bar{\mu}_2)\, V(\tau_0(\omega))}_{\text{failed transmission}} + \underbrace{(\bar{\omega}\mu_1 + \omega\mu_2)\, V(\tau_1(\omega))}_{\text{successful transmission}} \right)}^{\text{next step}}}_{\text{bandit}} \Bigg\}$$

The Bellman equation for average costs (no discounting) is similar.

Some Structural Results (Average Costs)

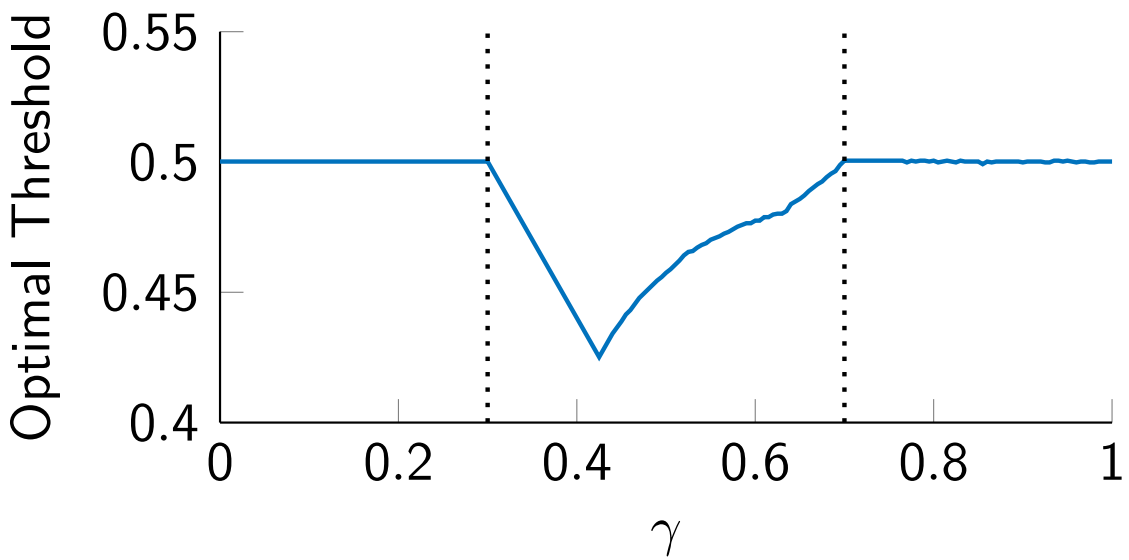
Numerical Observations

1. The optimal policy is a threshold policy: For $\omega < \omega^*$ choose 's', otherwise choose 'b'.
2. In comparison to the "myopic" threshold, $\omega^m = (\mu_s - \mu_1)/(\mu_2 - \mu_1)$, we have $\omega^* \leq \omega^m$

Some Structural Results (Average Costs)

Numerical Observations

1. The optimal policy is a threshold policy: For $\omega < \omega^*$ choose 's', otherwise choose 'b'.
2. In comparison to the "myopic" threshold, $\omega^m = (\mu_s - \mu_1)/(\mu_2 - \mu_1)$, we have $\omega^* \leq \omega^m$



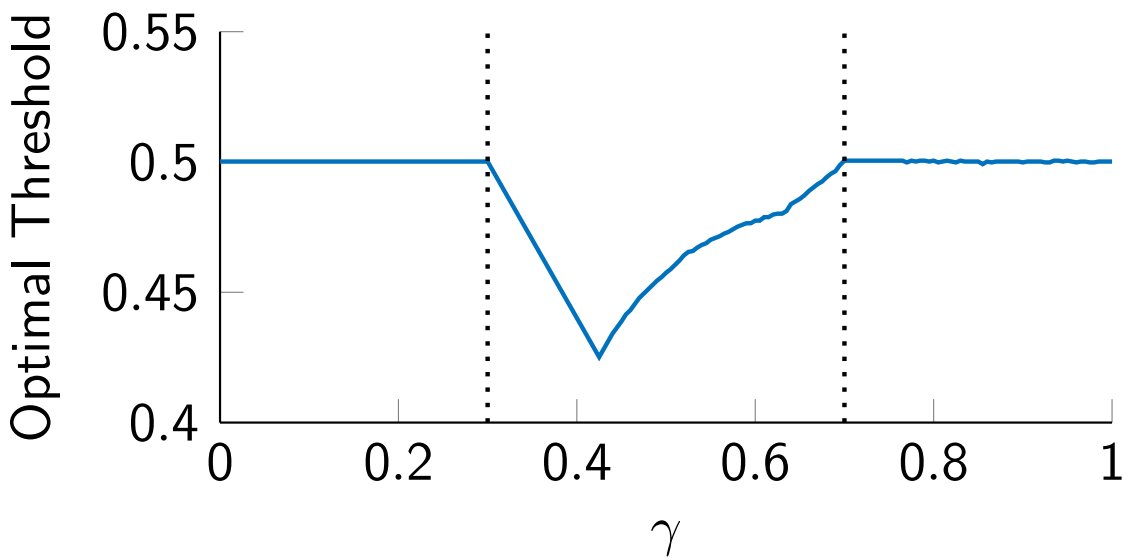
Optimal threshold values with $\mu_s = 0.5$, $\mu_1 = 0.2$, $\mu_2 = 0.8$, $\rho = 0.4$.

The attraction region of $\tau_i(\cdot)$ is marked by the vertical dotted lines.

Some Structural Results (Average Costs)

Numerical Observations

1. The optimal policy is a threshold policy: For $\omega < \omega^*$ choose 's', otherwise choose 'b'.
2. In comparison to the "myopic" threshold, $\omega^m = (\mu_s - \mu_1)/(\mu_2 - \mu_1)$, we have $\omega^* \leq \omega^m$

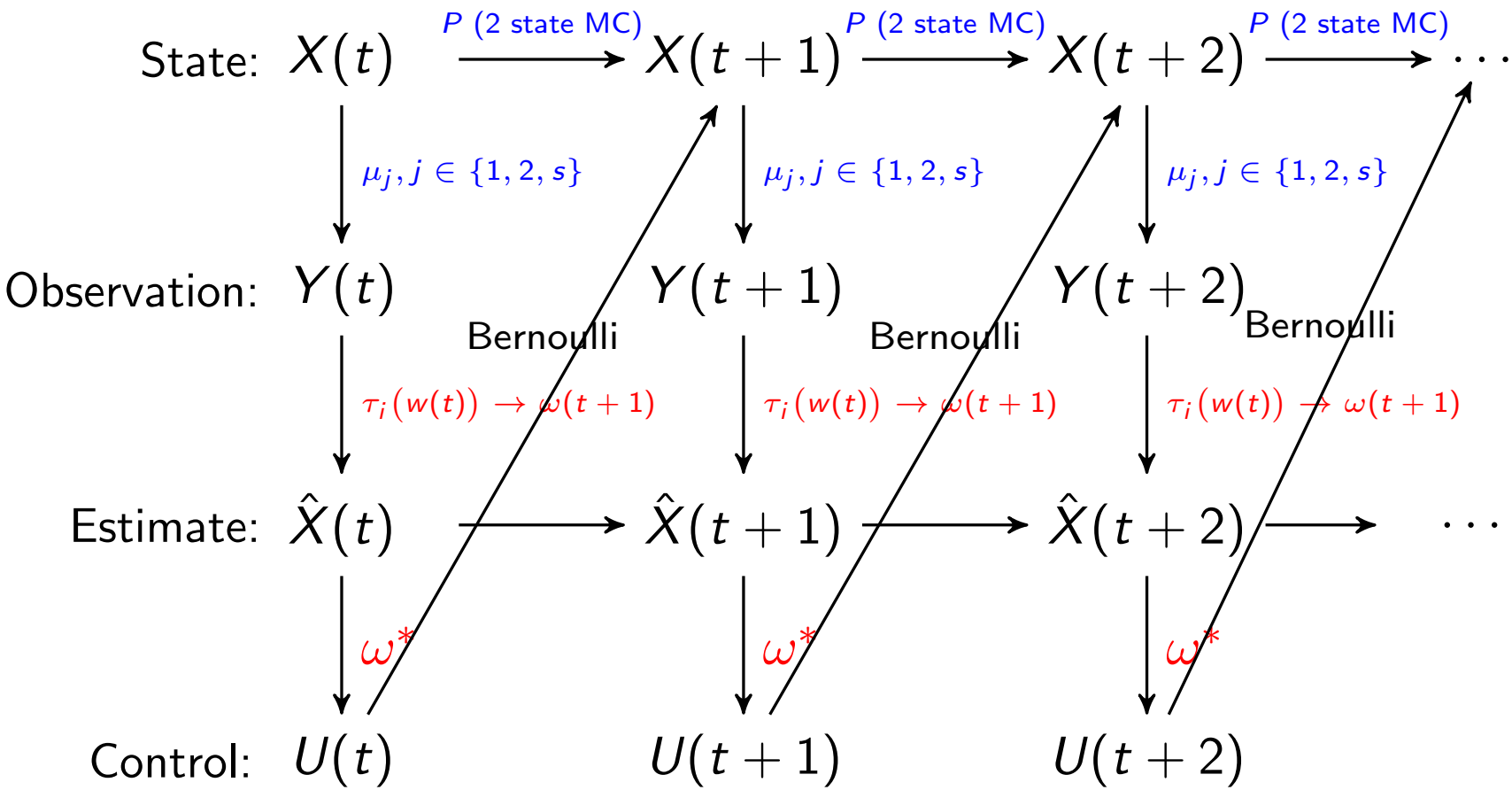


Optimal threshold values with $\mu_s = 0.5$, $\mu_1 = 0.2$, $\mu_2 = 0.8$, $\rho = 0.4$.

The attraction region of $\tau_i(\cdot)$ is marked by the vertical dotted lines.

Moral: Myopic is typically not the best. Need to take exploration into account in optimal decision.

Simple Server Selection Summary



Hope: Proving optimality of ω^* threshold (structural result)

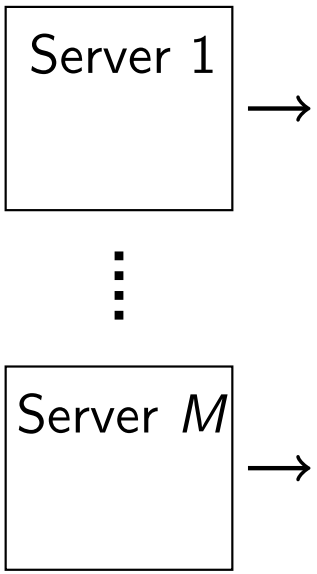
Less Hope: An explicit ω^*

Hope: Rougher structural results for the more general problem

Reward Observing Restless Bandits

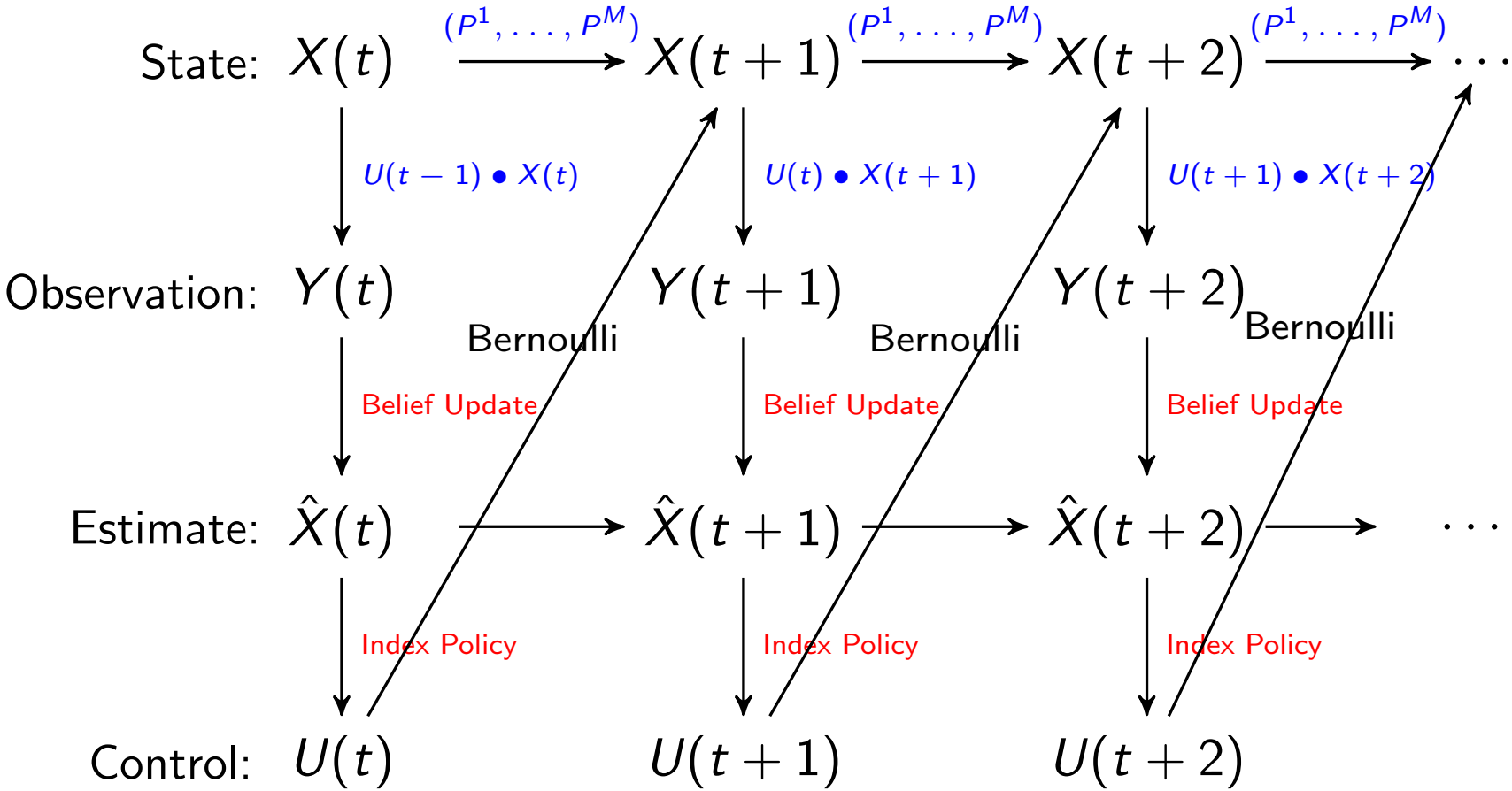
(based on some joint work with J. Kuhn)

Several (or Many) Servers to Choose From (No Queues)



- $U(t)$: choose $d < M$ servers, to maximise long term reward
- Server states, $X_i(t)$, evolve independently and are fully observed when the server is selected, otherwise not observed
- Easy to handle server state models are:
  - Two state Markov chains
  - Autoregressive processes of order 1
- Explicit (numerical) solution as a POMDP is hopeless
  $\Longrightarrow$ use index policies on the belief state (approximation)

Reward Observing Restless Bandits Summary



Index Policies

1. The belief state for each channel i is $\omega_i \in [0,1]$ for 2-state MC channels and $(\mu_i, \nu_i) \in \mathbb{R} \times \mathbb{R}_+$ for AR channels
2. For each channel set an index function $\mathcal{I}_i(\text{belief state}) \to \mathbb{R}$
3. Policy: $U(t)$ indicates the d channels with the highest index

Index Policies

1. The belief state for each channel i is $\omega_i \in [0,1]$ for 2-state MC channels and $(\mu_i, \nu_i) \in \mathbb{R} \times \mathbb{R}_+$ for AR channels
2. For each channel set an index function $\mathcal{I}_i(\text{belief state}) \to \mathbb{R}$
3. Policy: $U(t)$ indicates the d channels with the highest index

The Celebrated Gittins Index Result(s)

If $d = 1$ and channels freeze when not being selected, then an *optimal* policy is a specific index (Gittins et. al.).

Index Policies

1. The belief state for each channel i is $\omega_i \in [0,1]$ for 2-state MC channels and $(\mu_i, \nu_i) \in \mathbb{R} \times \mathbb{R}_+$ for AR channels
2. For each channel set an index function $\mathcal{I}_i(\text{belief state}) \to \mathbb{R}$
3. Policy: $U(t)$ indicates the d channels with the highest index

The Celebrated Gittins Index Result(s)

If $d = 1$ and channels freeze when not being selected, then an *optimal* policy is a specific index (Gittins et. al.).

Restless Bandits and the Whittle Index

The Reward Observing Restless Bandits problem is a specific case of the *Restless Bandits* problem of P. Whittle (1988).
A solution of a relaxed problem is an index policy. In certain cases as $M \to \infty$ the policy becomes optimal.

Exploration vs. Exploitation with Partially Observable AR(1) Arms

Julia Kuhn
Supervisors: Yoni Nazarathy (UQ) & Michel Mandjes (Universiteit van Amsterdam)

The University Of Queensland
AUSTRALIA

I. Model and Framework

A dynamic decision problem under uncertainty: We select k out of d restless *reward observing* one-armed bandits to play on, such as to maximize the expected total discounted or average reward. Rewards are collected and states are observed ONLY if an arm is played.

Should we collect new information or opt for the immediate payoff?

State processes are Gaussian AR(1),

$$X_i(t) = \varphi\, X_i(t-1) + \varepsilon_i(t),$$

where $\varphi \in (0,1)$ and $\varepsilon_i \sim$ i.i.d. $\mathcal{N}(0,\sigma^2)$. An application is channel selection in wireless networks.

Why is the Gaussian model so special?

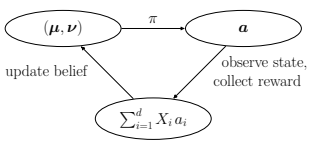
- The *belief states* $\big(\mu_i(t),\, \nu_i(t)\big)$, i.e. the means and variances conditioned on the available information, contain all relevant information available at time t.
- At the same time, $\mu_i(t)$ and $\nu_i(t)$ quantify the expected gain from exploiting an arm vs. the need for exploring it.

Updating the Belief States

A *policy* π maps the information available to actions $a_i(t) = 1$ ("play") or $a_i = 0$ ("do not play"), such that in total k out of d are played at every time t. With $Y_{\mu,\nu} \sim \mathcal{N}(\mu,\nu)$,

$$\big(\mu_i(t+1),\, \nu_i(t+1)\big) = \begin{cases} \big(\varphi\,\mu_i(t),\, \varphi^2\nu_i(t) + \sigma^2\big), & a_i(t) = 0, \\ \big(\varphi\, Y_{\mu_i(t),\,\nu_i(t)},\, \sigma^2\big), & a_i(t) = 1. \end{cases}$$

Chain of Actions



References

1. K. Avrachenkov, L. Cottatellucci, and L. Maggi (2012). Slow Fading Channel Selection: A Restless Multi-armed Bandit Formulation. *ISWCS*, pp. 1083–1087.
2. J. Kuhn, M. Mandjes and Y. Nazarathy (2014). Exploration vs. Exploitation with Partially Observable Gaussian Autoregressive Arms. *Submitted.*
3. J. Gittins, K. Glazebrook and R. Weber (2011). *Multi-armed Bandit Allocation Indices*, 2nd Ed., John Wiley & Sons.
4. P. Whittle (1988). Restless bandits: Activity Allocation in a Changing World. *Journal of Applied Probability*, 25, pp. 287–298.

II. Index Policies

An index policy is of the form

$$\pi_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \underset{\boldsymbol{a}:\sum_{i=1}^d a_i = k}{\arg\max} \left\{ \sum_{i=1}^{d} \gamma\,(\mu_i, \nu_i)\, a_i \right\}$$

The *index function* γ maps the belief state of each arm to some priority index.

Myopic	$\gamma^M(\mu,\nu) = \mu$
Parametric	$\gamma^\theta(\mu,\nu) = \mu + \theta\nu, \quad \theta > 0$
Whittle	$\gamma^W(\mu,\nu) = \inf\{\lambda \,|\, \pi_{\text{opt}}(\mu,\nu) = 0\}$

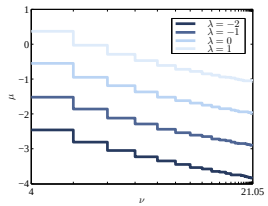
Here π_{opt} is the optimal policy for a *one-armed bandit problem with subsidy*, where the decision maker observes and collects the reward when playing, and obtains a subsidy λ otherwise.

Motivated by the many-arms asymptotic behaviour of the system.

III. Whittle Index: Structural Results

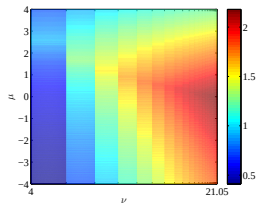
The Whittle index policy has been found to be asymptotically optimal in many cases (although no such result is known for our model) but no closed-form expression is known. The associated optimal value function can in principle be found using dynamic programming techniques. We can further prove the following.

The optimal policy for the one-armed bandit problem with subsidy is a *threshold policy*.



Switching curves: above the curve the optimal action is "play", below "do not play".
$\beta = 0.8,\ \varphi = 0.9,\ \sigma = 2.$

The Whittle index $\gamma^W(\mu,\nu)$ is monotone non-decreasing in μ and ν, and generally not constant.



Difference of Whittle and myopic index:
$\gamma^W(\mu,\nu) - \mu.$
$\beta = 0.8,\ \varphi = 0.9,\ \sigma = 2.$

IV. Parametric Index: Many-Arms Asymptotic Behaviour

1. Intuitively, as $d \to \infty$, $k_d/d \to \rho$, in the long-run the system approaches an equilibrium at which the proportion of arms associated with a certain belief state remains fixed. Then the action chosen for a certain arm is independent of the current belief state of any other arm, as there is always the same proportion of arms associated with a certain belief state in the system.

2. We explicitly identify a measure-valued recursion that describes the many-arms behaviour of the system at equilibrium. Namely, the limiting proportion of arms that have been observed h time steps ago and whose conditional mean falls in $(-\infty, x]$ can be modeled as

$$m_h\big(x,\, t+1\big) = \begin{cases} \sum_{h=0}^{\infty} \int_{\ell_h^*(t)}^{\infty} \Phi_{z,\,\nu^{(h)}}\left(\frac{x}{\varphi}\right) m_h(dz, t), & h = 0, \\ m_{h-1}\Big(\min\big\{\frac{x}{\varphi},\, \ell_{h-1}^*(t)\big\},\, t\Big), & h \geq 1, \end{cases}$$

where $\ell_h^*(t) := \ell^*(t) - \theta\nu^{(h)}(t)$ with $\ell^*(t)$ defined by

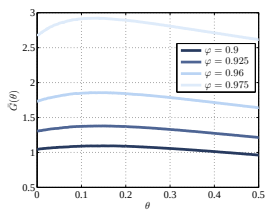
$$\ell^*(t) = \sup\left\{ \ell \,\Big|\, \sum_{h=0}^{\infty} m_h\Big(\big\{\mu \,|\, \mu + \theta\nu^{(h)} \in [\ell,\infty)\big\},\, t\Big) = \rho \right\}.$$

Thus, $\ell_h^*(t)$ is a threshold such that at time t the parametric policy activates all arms that are of age h and have conditional mean $\mu(t) \geq \ell_h^*(t)$.

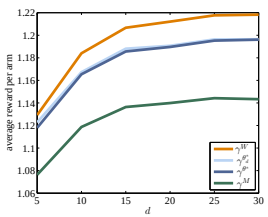
3. Based on **1.** and **2.** we conjecture that the measure-valued dynamical system at equilibrium is directly related to a one-armed process where the arm is activated whenever the index exceeds a particular threshold $\overline{\ell}$, namely $\overline{\ell} = \ell^*$.

Algorithm for Performance Evaluation

1. For large T determine $\overline{\ell}$ such that $T^{-1}\sum_{t=0}^{T} a_i(t) = \rho$ is achieved for a parametric index policy applied to the one-armed process.
2. Use the sample path of Step 1 to obtain an estimate $\overline{G}$ for the expected average reward of the one-armed system.
3. Output $\overline{G}_d := d\,\overline{G}$ as an approximation of the expected average reward of the multiarmed system with d arms.



Expected average reward $\overline{G}(\theta)$ computed by the algorithm as a function of θ. $\sigma = 2$, $\varphi \in \{0.9, 0.925, 0.95, 0.975\}$, $\rho = 0.4$, $T = 2 \times 10^6$.



Comparison of average rewards achieved per arm. θ is found by optimizing (i) the problem with d arms (θ_d^*), and (ii) the one-armed problem (θ^*). $\varphi = 0.9$, $\sigma = 2$, $\rho = 0.4$, $T = 10^5$.

References

- P. Whittle, "*Optimization Over Time: Dynamic Programming and Stochastic Control*", John Wiley & Sons, 1982.
- P. Antsaklis and A. Michel, "*A Linear Systems Primer*", Birkhauser Boston, 2007.
- P. Whittle, "*Restless Bandits: Activity allocation in a changing world.*", Journal of Applied Probability, 1988.
- Y. Nazarathy, T. Taimre, A. Asanjarani, J. Kuhn, B. Patch and A. Vuorinen, "*The Challenge of Stabilizing Control for Queueing Systems with Unobservable Server States*", Proceedings of AUCC Confrence, accepted for publication, 2015.
- J. Kuhn, Y. Nazarathy "*Wireless Channel Selection with Reward-Observing Restless Multi-armed Bandits*", Chapter to appear in "Markov Decision Processes in Practice", Editors: R. Boucherie and N. van Dijk.
- J. Kuhn, M. Mandjes and Y. Nazarathy "*Exploration vs. Exploitation with Partially Observable Gaussian Autoregressive Arms*", Proceedings of the Valuetools Conference, 2014.
- Y. Nazarathy, L. Rojas-Nandayapa and T. Salisbury "*Non-existence of Stabilizing Policies for the Critical Push-Pull Network and Generalizations*", Operations Research Letters, 2013.