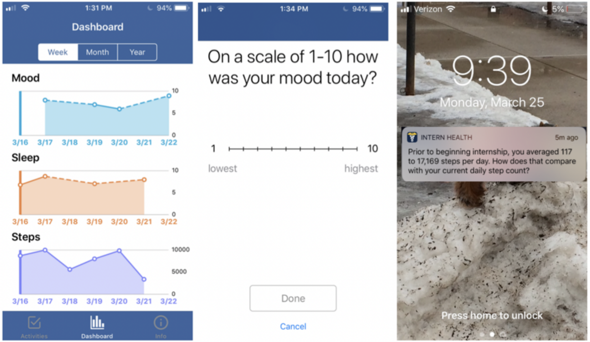
Doubly Inhomogeneous Reinforcement Learning

Chengchun Shi
Assistant Professor of Data Science
London School of Economics and Political Science
Joint Work with Liyuan Hu, Mengbing Li, Zhenke Wu and Piotr Fryzlewicz

Developing AI with Reinforcement Learning

Mobile Health (mHealth)

- **Data**: Intern Health Study (NeCamp et al., 2020)
- **Subject**: First-year medical interns working in stressful environments (e.g., long work hours and sleep deprivation)
- **Objective**: Promote physical and mental well-being
- **Intervention**: Determine whether to send certain text message to a subject



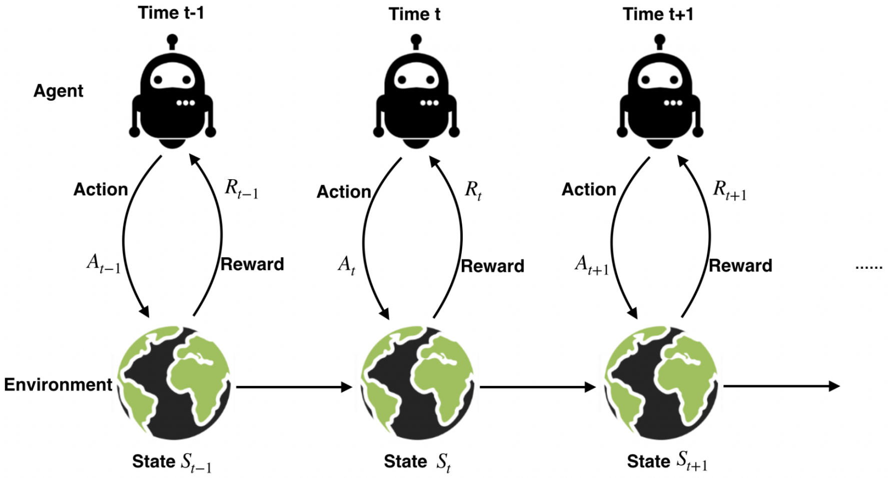
(i) App Dashboard (ii) Mood EMA (iii) Notifications

Intern Health Study

Table 1. Examples of 6 different groups of notifications.

Notification groups	Life insight	Tip
Mood	Your mood has ranges from 7 to 9 over the past 2 weeks. The average intern's daily mood goes down by 7.5% after intern year begins.	Treat yourself to your favorite meal. You've earned it!
Activity	Prior to beginning internship, you averaged 117 to 17,169 steps per day. How does that compare with your current daily step count?	Exercising releases endorphins which may improve mood. Staying fit and healthy can help increase your energy level.
Sleep	The average nightly sleep duration for an intern is 6 hours 42 minutes. Your average since starting internship is 7 hours 47 minutes.	Try to get 6 to 8 hours of sleep each night if possible. Notice how even small increases in sleep may help you to function at peak capacity & better manage the stresses of internship.

Sequential Decision Making



Objective: find an optimal policy that maximizes the cumulative reward

Reinforcement Learning

- **RL algorithms**: trust region policy optimization (Schulman et al., 2015), deep Q-network (DQN, Mnih et al., 2015), asynchronous advantage actor-critic (Minh et al., 2016), quantile regression DQN (Dabney et al., 2018).
- **Foundations** of RL:
  - **Markov decision process** (MDP, Puterman, 1994): ensures the optimal policy is *stationary* over time and *homogeneous* across subjects
  - **Markov assumption** (MA): Within each data trajectory, conditional on the present (e.g., S_t, A_t), the future (R_t, S_{t+1}) and the past data history are independent
  - **Global stationarity assumption** (GSA): Within each data trajectory, the Markov transition kernel is stationary over time
  - **Global homogeneity assumption** (GHA): At each time, all data trajectories share the same Markov transition kernel

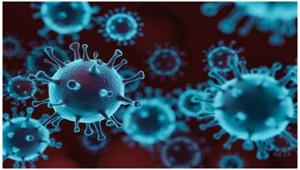
Reinforcement Learning

- **RL algorithms**: trust region policy optimization (Schulman et al., 2015), deep Q-network (DQN, Mnih et al., 2015), asynchronous advantage actor-critic (Minh et al., 2016), quantile regression DQN (Dabney et al., 2018).
- **Foundations** of RL:
  - **Markov decision process** (MDP, Puterman, 1994): ensures the optimal policy is *stationary* over time and *homogeneous* across subjects
  - **Markov assumption** (MA): Within each data trajectory, conditional on the present (e.g., $\boldsymbol{S_t}$, $\boldsymbol{A_t}$), the future ($\boldsymbol{R_t}$, $\boldsymbol{S_{t+1}}$) and the past data history are independent
  - **Global stationarity assumption** (GSA): Within each data trajectory, the Markov transition kernel is stationary over time
  - **Global homogeneity assumption** (GHA): At each time, all data trajectories share the same Markov transition kernel

Temporal Non-stationarity & Subject Heterogeneity

(a) Mobile Health

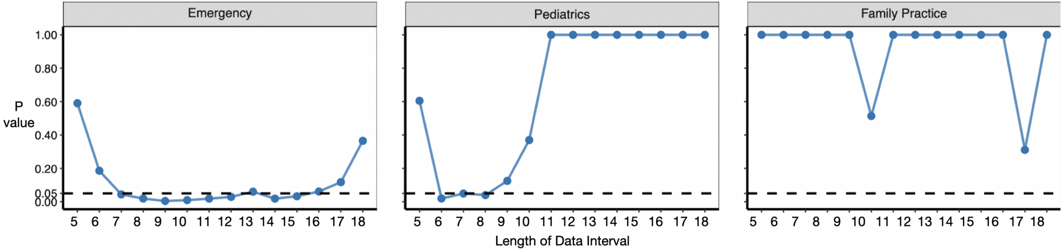
(b) Ridesharing



(c) Infectious Disease Control

- **Violation of GSA**
  (a) treatment effects decay over time
  (b) weekday-weekend differences, peak and off-peak differences
  (c) COVID mutations, development of vaccines
- **Violation of GHA**
  (a) patient's heterogeneity toward treatment
  (b) supply (no. of drivers) & demand (no. of call orders) differ across cities
  (c) population density & health insurance system differ across regions

Intern Health Study (Revisit)



- Cluster medical interns according to their specialties
- Test the stationarity assumption over a sequence of data intervals
- A significant p-value indicates the existence of a change point

Double Inhomogeneity

We study RL in **doubly inhomogeneous** environments (e.g., Markov transition kernel change over time and across subjects)

Table: Forms of the Optimal Policy in Different Environments.

GSA ✓ GHA ✓	GSA ✓ GHA ✗	GSA ✗ GHA ✓	GSA ✗ GHA ✗
doubly homogeneous	stationary	homogeneous	subject-specific history-dependent

Configurations of Double Inhomogeneity

- To illustrate double inhomogeneity, consider two subjects with a single change point

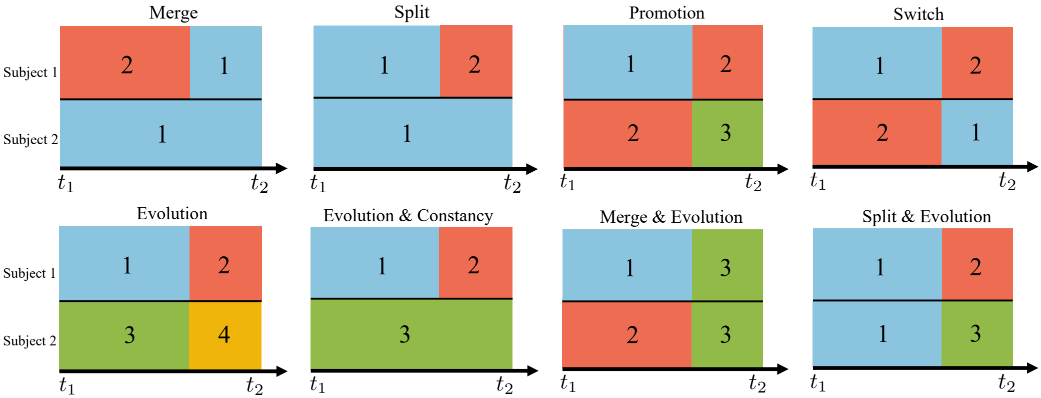


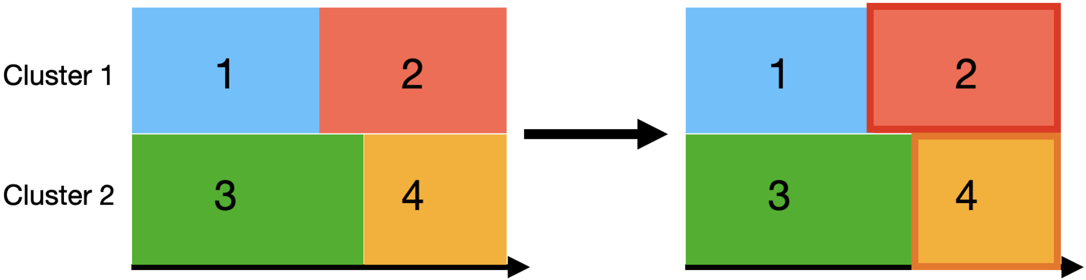
Figure: Basic building blocks with two subjects (one in each row) and a single change point. Different transition dynamics are represented by distinct colors/numbers.

Data, Assumptions and Objective

- Data: N trajectories, T time points per trajectory.
- Question: how to learn an optimal policy for these subjects at time T?
- Challenge: **borrow information** in the presence of double inhomogeneity
- Our assumptions:
  1. **Local Stationarity at the Endpoint** (LSE): For each subject i, there exists some $\tau_i > 0$ such that the Markov transition kernel is a constant function of t for any $T - \tau_i \leq t \leq T$.
  2. **Local Homogeneity at the Endpoint** (LHE): There exists a finite number K of disjoint subject clusters $\cup_{k=1}^{K} \mathcal{C}_k$, where $\mathcal{C}_k \subseteq \{1, ..., N\}$, such that within each cluster $\mathcal{C}_k$, the Markov transition kernel at time T is constant over different subjects
- Objective: determine the **best data rectangle** that display similar dynamics over time and subjects for effective policy learning

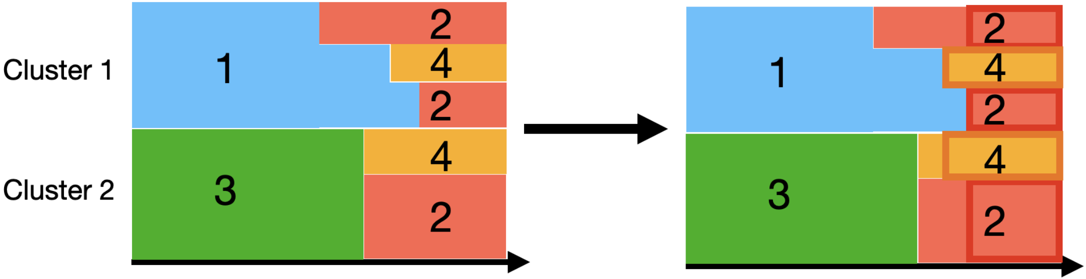
Best Data Rectangle

A simple example ...

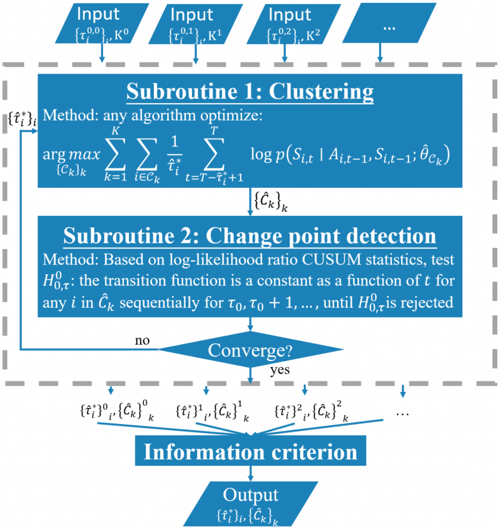


Best Data Rectangle (Cont'd)

A slightly more complicated example ...



Method


Input $\{\tau_i^{0,0}\}_i, K^0$
Input $\{\tau_i^{0,1}\}_i, K^1$
Input $\{\tau_i^{0,2}\}_i, K^2$
...
Subroutine 1: Clustering
Method: any algorithm optimize:
$\arg\max_{\{\mathcal{C}_k\}_k} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \frac{1}{\hat{\tau}_i^*} \sum_{t=T-\hat{\tau}_i^*+1}^{T} \log p(S_{i,t} \mid A_{i,t-1}, S_{i,t-1}; \hat{\theta}_{\mathcal{C}_k})$
$\{\hat{\tau}_i^*\}_i$
$\{\hat{C}_k\}_k$
Subroutine 2: Change point detection
Method: Based on log-likelihood ratio CUSUM statistics, test $H_{0,\tau}^0$: the transition function is a constant as a function of t for any i in $\hat{C}_k$ sequentially for $\tau_0, \tau_0+1, \dots,$ until $H_{0,\tau}^0$ is rejected
no
Converge?
yes
$\{\hat{\tau}_i^*\}^0{}_i, \{\hat{C}_k\}^0{}_k$
$\{\hat{\tau}_i^*\}^1{}_i, \{\hat{C}_k\}^1{}_k$
$\{\hat{\tau}_i^*\}^2{}_i, \{\hat{C}_k\}^2{}_k$
...
Information criterion
Output $\{\hat{\tau}_i^*\}_i, \{\hat{C}_k\}_k$

Theory

Table 2: Rate of convergence when N and T have different divergence properties. The "CP error" refers to the change point detection error and "non-negligible" means that the error does not decay to zero as $N \to \infty$.

Iteration		$T \to \infty$ $N \to \infty$	$T \to \infty$ N fixed	T fixed $N \to \infty$
1^{st}	clustering error	0	0	non-negligible
	CP error	0	$O_p\left(\frac{\log^2(NT)}{NTs_{cp}^2}\right)$	non-negligible
2^{nd}	clustering error	0	0	non-negligible
	CP error	0	$O_p\left(\frac{\log^2(NT)}{NTs_{cp}^2}\right)$	non-negligible
...		...	...	...

- Only require the **overestimation** error of each initial τ_i to satisfy certain rate. No assumption is imposed on their **underestimation** error.
- Detect **weaker signals** and have **faster convergence rates** compared to applying the clustering algorithm per time or the CP detection algorithm per subject

Simulation

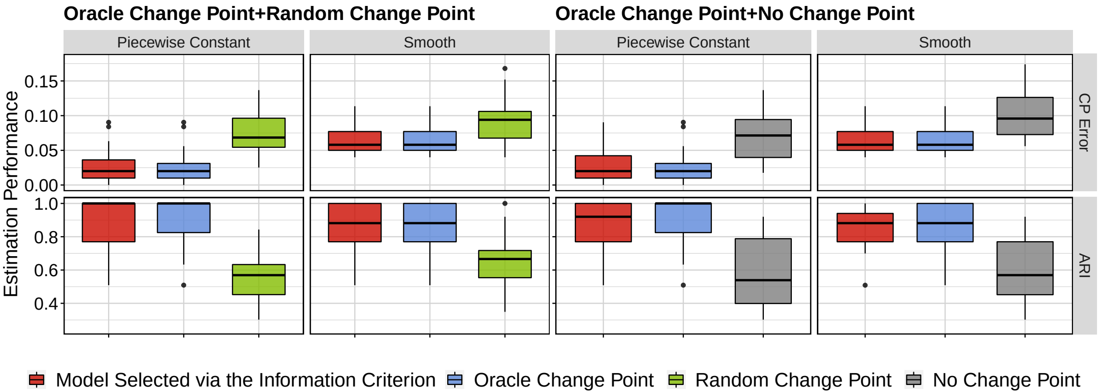


Figure: Average CP error and ARI with different initial change point locations are chosen by the information criterion.

Simulation (Cont'd)

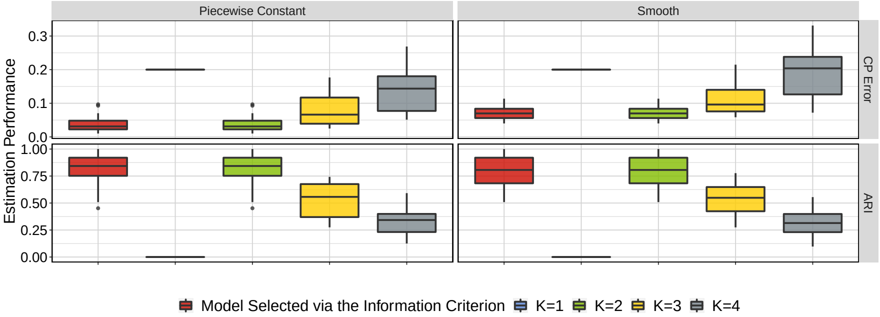


Figure: Average performance in offline estimation with different number of clusters ($K = 1, 2, 3, 4$) and the results chosen by the information criterion.

Simulation (Cont'd)

- **Online value evaluation**: recursively apply the proposed algorithm to update the estimated optimal policy and use this policy for action generation
- **Competing policies**: oracle, doubly homogeneous (DH), homogeneous, stationary

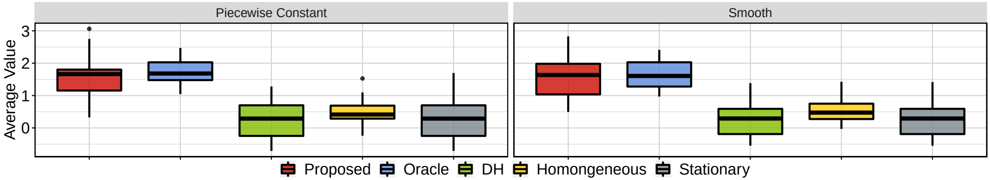


Figure: Boxplot of the expected returns under the proposed policy and other baseline policies that either ignore non-stationarity or heterogeneity.

Thank You!

☺Papers and softwares can be found on my personal website

`callmespring.github.io`