Reinforcement Learning in Possibly Nonstationary Environment

Chengchun Shi

Assistant Professor of Data Science

London School of Economics and Political Science

Joint work with Mengbing Li, Zhenke Wu and Piotr Fryzlewicz

Developing AI with Reinforcement Learning

In this talk, we will focus on ...

- Reinforcement learning in **offline real-world applications** (e.g., mobile health, ridesharing).
  - Most works consider developing RL algorithms in games (online)

(a) Mobile Health

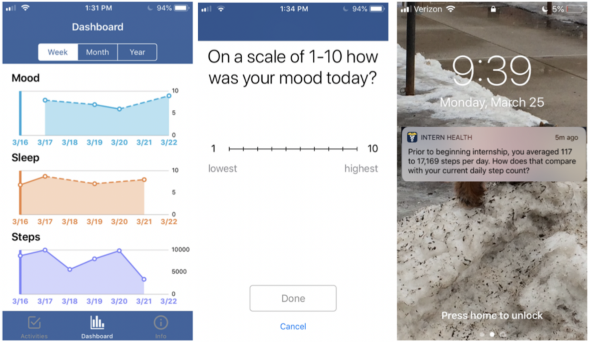
(b) Ridesharing

(c) Games

- **Statistical inference** in reinforcement learning
  - Is statistical inference useful in reinforcement leaning?

Intern Health Study (IHS)

- **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 well-being
- **Intervention**: Determine whether to send certain text message to a subject



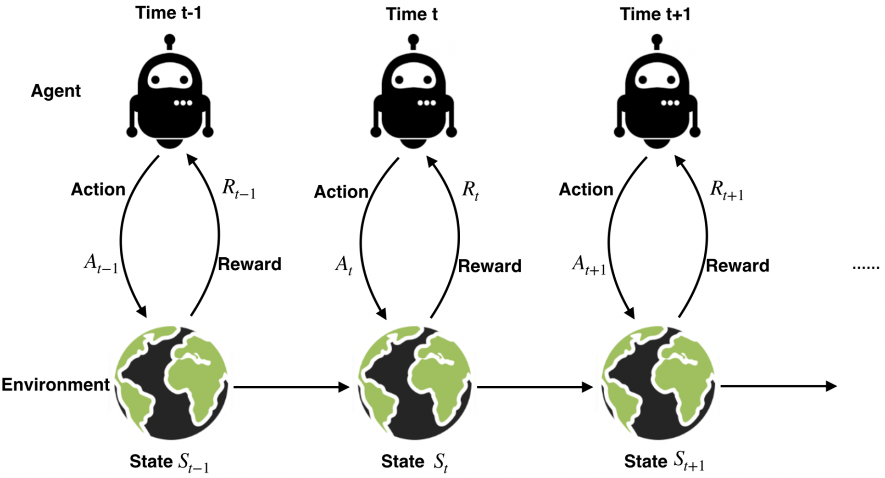
(i) App Dashboard (ii) Mood EMA (iii) Notifications

Intern Health Study (Cont'd)

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

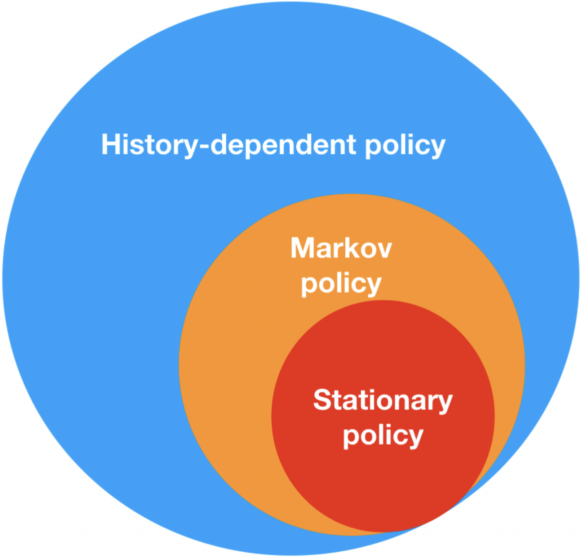
The Agent's Policy

- The agent implements a **mapping** π_t from the observed data to a probability distribution over actions at each time step
- The collection of these mappings $\pi = \{\pi_t\}_t$ is called **the agent's policy**:

$$\pi_t(a|\bar{s}) = \Pr(A_t = a|\bar{S}_t = \bar{s}),$$

  where $\bar{S}_t = (S_t, R_{t-1}, A_{t-1}, S_{t-1}, \cdots, R_0, A_0, S_0)$ is the set of **observed data history** up to time t.
- **History-Dependent** Policy: π_t depends on $\bar{S}_t$.
- **Markov** Policy: π_t depends on $\bar{S}_t$ only through S_t.
- **Stationary** Policy: π is Markov & π_t is **homogeneous** in t, i.e., $\pi_0 = \pi_1 = \cdots$.

The Agent's Policy (Cont'd)



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*, and is *not* history-dependent.
  - **Markov assumption**: conditional on the present (e.g., S_t, A_t), the future (R_t, S_{t+1}) and the past data history are independent
  - **Stationarity assumption**: the Markov transition kernel, e.g., the conditional distribution of (R_t, S_{t+1}) given ($S_t = s$, $A_t = a$) is stationary over time

Stationarity Assumption

- Stationarity assumption is likely to hold in many **OpenAI Gym** environments
- However, it can be violated in the **real world** environment
- Treatment effects can be **nonstationary**
    - COVID vaccine effectiveness decays over time
    - The treatment effect of activity suggestions may transition from positive to negative
- Environments can be **nonstationary**
    - COVID mutations, invention of vaccines
    - In the context of mobile-delivered prompts, the longer a person is under intervention, the more they may habituate to the prompts or become overburdened
- Without stationarity, the optimal policy is **nonstationary** as well
- Crucial for policy maker to take nonstationarity into account

Models with/without SA

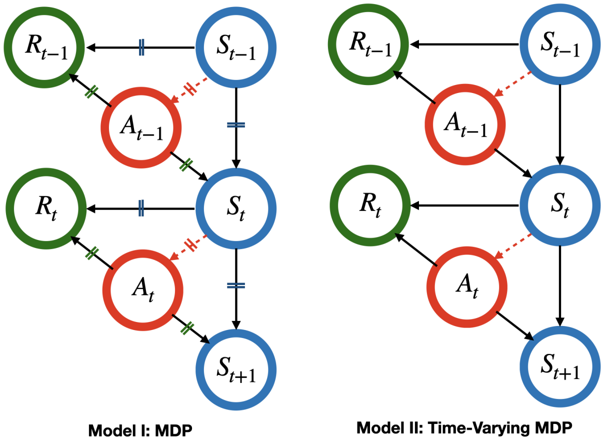


Figure: Causal diagrams for MDPs and TMDPs. Solid lines represent the causal relationships. Dashed lines indicate the information needed to implement the optimal policy. The parallel sign $\|$ indicates that the conditional probability function given parent nodes is equal.

Challenges

- When the optimal policy is **nonstationary**, using all data is not reasonable
- Natural to use **more recent observations** for policy optimisation
- **Challenging** to select the **best data "segment"**
  - Including too many past observations yields a suboptimal policy
  - Using only a few recent observations results in a very noisy policy

Contributions

- **Methodologically**
  - **First** work on developing consistent test for stationarity in offline RL
    - The test procedure is "**model-free**" (target on the optimal Q-function Q^{opt})
    - Null hypothesis $\boldsymbol{\mathcal{H}_0}$: Q^{opt} is stationary over time
    - Alternative hypothesis $\boldsymbol{\mathcal{H}_1}$: Q^{opt} varies over time
  - Sequentially apply the test for selecting the **best data "segment"**
- **Empirically**
  - Identify a **better** policy compared to existing RL algorithms in IHS
- **Theoretically**
  - prove our test has good **size** and **power** properties under a **bidirectional** asymptotic framework

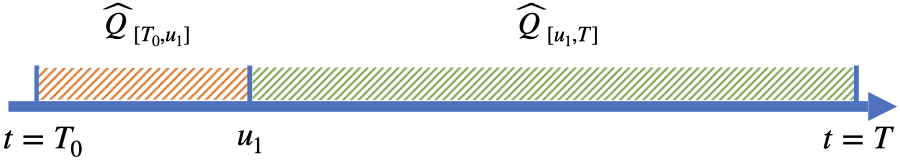
Method: Test Statistics

Some key components of the test statistic:

- Model the optimal Q-function via the **sieve** method
  - Ensure the estimator has a tractable limiting distribution
  - Increase the number of sieves to reduce the bias resulting from model misspecification
- Construct **CUSUM**-type test statistics for change detection (detailed later)
  - Widely used in the time series literature
- Obtain critical values using **multiplier bootstrap**
  - Q-estimator is asymptotically normal
  - Test statistic is a complicated function of several Q-estimators
  - Bootstrapped statistic is a function of simulated random normal errors
  - Approximate critical values via the quantile of the bootstrapped statistic

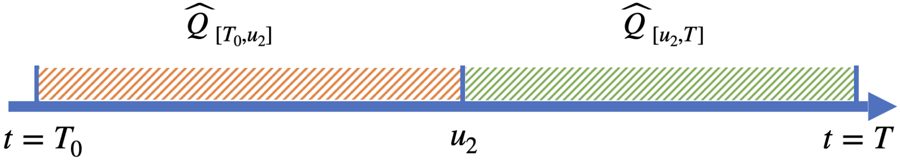
Method: Test Statistics (Cont'd)

- A **CUSUM**-type test statistic
  - Select a set of candidate change point locations $\boldsymbol{u} \in [\boldsymbol{T_0}, \boldsymbol{T}]$
  - For each $\boldsymbol{u}$, estimate two Q-functions $\widehat{\boldsymbol{Q}}_{[\boldsymbol{T_0},\boldsymbol{u}]}$ and $\widehat{\boldsymbol{Q}}_{[\boldsymbol{u},\boldsymbol{T}]}$
  - Construct the test based on their maximal difference



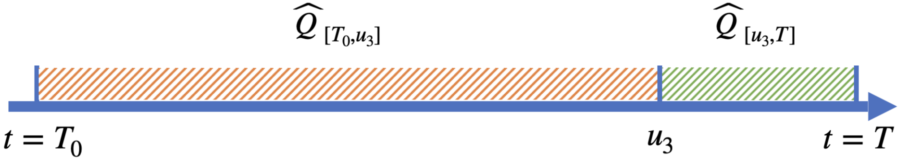
Method: Test Statistics (Cont'd)

- A **CUSUM**-type test statistic
  - Select a set of candidate change point locations $\boldsymbol{u} \in [\boldsymbol{T_0}, \boldsymbol{T}]$
  - For each $\boldsymbol{u}$, estimate two Q-functions $\widehat{\boldsymbol{Q}}_{[\boldsymbol{T_0}, \boldsymbol{u}]}$ and $\widehat{\boldsymbol{Q}}_{[\boldsymbol{u}, \boldsymbol{T}]}$
  - Construct the test based on their maximal difference



Method: Test Statistics (Cont'd)

- A **CUSUM**-type test statistic
  - Select a set of candidate change point locations $\boldsymbol{u} \in [\boldsymbol{T_0}, \boldsymbol{T}]$
  - For each $\boldsymbol{u}$, estimate two Q-functions $\widehat{\boldsymbol{Q}}_{[\boldsymbol{T_0},\boldsymbol{u}]}$ and $\widehat{\boldsymbol{Q}}_{[\boldsymbol{u},\boldsymbol{T}]}$
  - Construct the test based on their maximal difference



Method: Test Statistics (Cont'd)

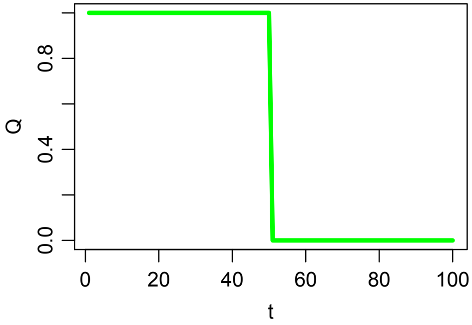
- Standard CUSUM-statistics that focuses on the difference in the **mean**
- We focus on the difference in Q which is a **function** of the state-action pair
- Need to aggregate the maximal difference

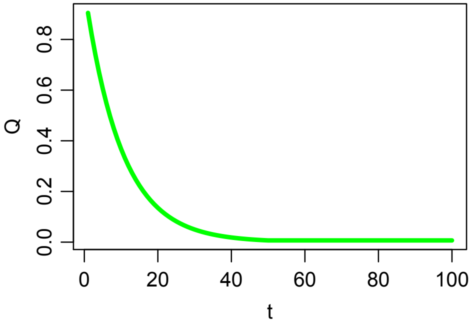
$$\Delta(a,s) = \max_{u} \sqrt{\frac{(T-u)(u-T_0)}{(T-T_0)}} |\widehat{Q}_{[T_0,u]}(a,s) - \widehat{Q}_{[u,T]}(a,s)| \tag{1}$$

  over different state-action pair
- Three proposed test statistics
  1. ℓ_1**-type**: aggregate $\Delta(a,s)$ over the empirical data distribution
  2. **maximum-type**: $\max_{a,s} \Delta(a,s)$
  3. **normalized maximum** (widely used in econ): $\max_{a,s} \widehat{\sigma}^{-1}(a,s)\Delta(a,s)$
- Bootstrapped statistic: replace $\widehat{Q}$ in (1) with simulated normal errors

Method: Test Statistics (Cont'd)

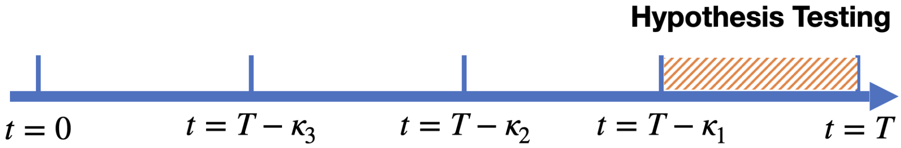
The test is able to detect both **abrupt** and **smooth** changepoints





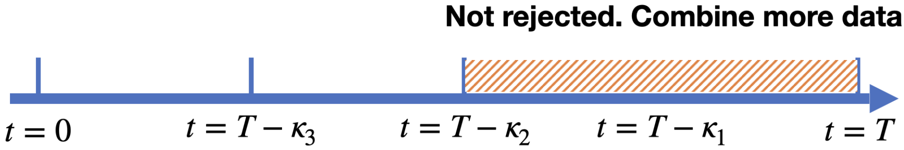
Method: Sequential Procedure

- Sequentially apply the test for selecting the **best data "segment"**
  - Sequentially test whether $\mathcal{H}_0$ holds on the data interval $[T-\kappa, T]$ for $\kappa_1 < \kappa_2 < \kappa_3 < \cdots$
  - Suppose $\mathcal{H}_0$ is first rejected at some $\kappa = \kappa_{j_0}$
  - Use the data subset within the interval $[T-\kappa_{j_0-1}, T]$ for policy optimisation



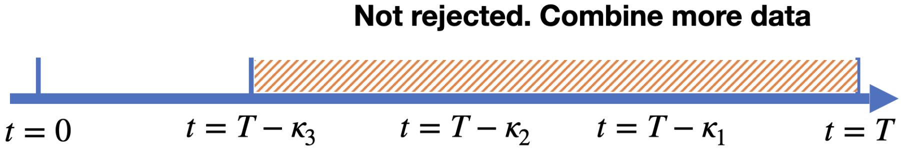
Method: Sequential Procedure (Cont'd)

- Sequentially apply the test for selecting the **best data "segment"**
  - Sequentially test whether $\mathcal{H}_0$ holds on the data interval $[T - \kappa, T]$ for $\kappa_1 < \kappa_2 < \kappa_3 < \cdots$
  - Suppose $\mathcal{H}_0$ is first rejected at some $\kappa = \kappa_{j_0}$
  - Use the data subset within the interval $[T - \kappa_{j_0 - 1}, T]$ for policy optimisation



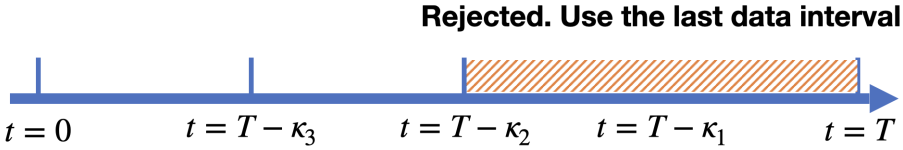
Method: Sequential Procedure (Cont'd)

- Sequentially apply the test for selecting the **best data "segment"**
  - Sequentially test whether $\mathcal{H}_0$ holds on the data interval $[T - \kappa, T]$ for $\kappa_1 < \kappa_2 < \kappa_3 < \cdots$
  - Suppose $\mathcal{H}_0$ is first rejected at some $\kappa = \kappa_{j_0}$
  - Use the data subset within the interval $[T - \kappa_{j_0 - 1}, T]$ for policy optimisation



Method: Sequential Procedure (Cont'd)

- Sequentially apply the test for selecting the **best data "segment"**
  - Sequentially test whether $\mathcal{H}_0$ holds on the data interval $[T-\kappa, T]$ for $\kappa_1 < \kappa_2 < \kappa_3 < \cdots$
  - Suppose $\mathcal{H}_0$ is first rejected at some $\kappa = \kappa_{j_0}$
  - Use the data subset within the interval $[T-\kappa_{j_0-1}, T]$ for policy optimisation

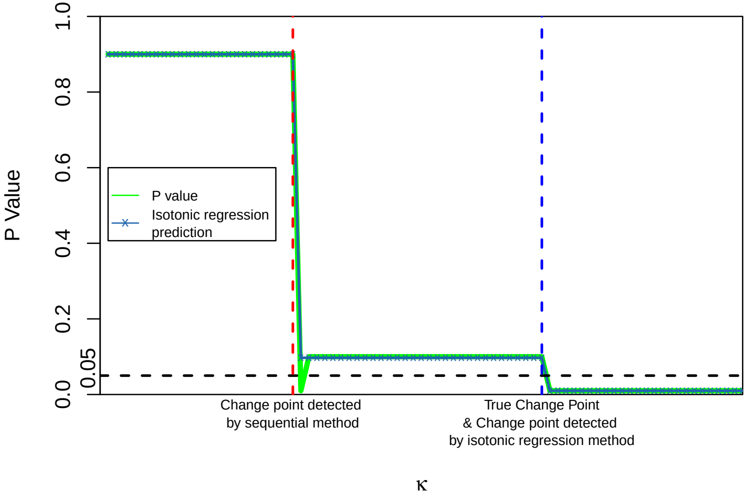


Method: Sequential Procedure (Cont'd)

- Our proposal: an improved version of the sequential procedure based on **isotonic regression**
- Main idea: When the data interval consists of a **single** change point, those significant p-values are **monotonic** over time
- Method:
  1. Sequentially test whether $\mathcal{H}_0$ holds on the data interval $[T - \kappa, T]$ for $\kappa_1 < \kappa_2 < \kappa_3 < \cdots$ and compute the p-value
  2. Apply isotonic regression to fit these p-values
  3. Suppose $\mathcal{H}_0$ is first rejected at some $\kappa = \kappa_{j_0}$, based on the fitted p-value
  4. Use the data subset within the interval $[T - \kappa_{j_0 - 1}, T]$ for policy optimisation
- The single-change-point assumption can be relaxed (see the data example)

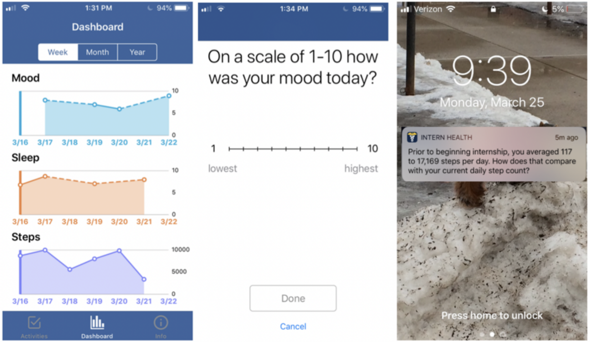
Method: Sequential Procedure (Cont'd)

The advantage of using isotonic regression



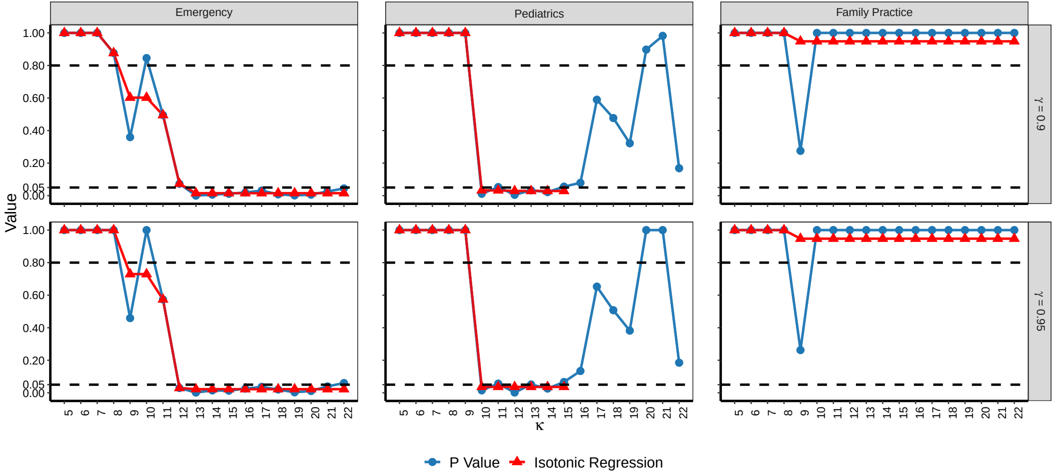
Application: Intern Health Study

- **Subject**: First-year medical interns
- **Objective**: Develop treatment policy to determine whether to send certain text messages to interns to improve their health
- S_t: Interns' **mood scores**, **sleep hours** and **step counts**
- A_t: Send **text notifications** or not
- R_t: **Step counts**



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

Application: Intern Health Study (Cont'd)



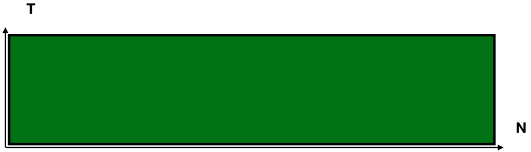
Application: Intern Health Study (Cont'd)

# Change points	Specialty	Method	Value	
			$\gamma = 0.9$	$\gamma = 0.95$
1	Emergency	Proposed	8073.27	8003.38
		Overall	7902.39	7794.77
		Behavior	7823.75	7777.32
≥ 2	Pediatrics	Proposed	7783.86	7762.81
		Overall	7680.04	7686.46
		Behavior	7730.98	7721.29
0	Family Practice	Proposed	8087.15	8072.78
		Overall	8087.15	8072.78
		Behavior	7967.67	7957.24

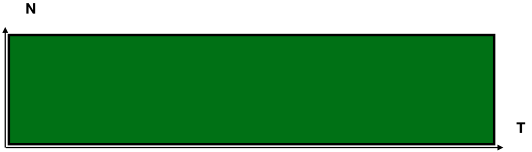
- Mean value is the weekly average step counts per day
- The proposed method improves mean value by 50 – 250 steps, compared to the behavior policy

Bidirectional Theory

- N the number of trajectories
- T the number of decision points per trajectory
- **bidirectional asymptotics**: a framework allows either N or $T \to \infty$
- large N, small T (Intern Health Study)



- small N, large T (OhioT1DM dataset)



- large N, large T (games)

Bidirectional Theory (Cont'd)

Theorem (Informal Statement)

Under certain conditions, as either $\boldsymbol{N}$ or $\boldsymbol{T}$ diverges to infinity

1. *Our test controls the type-I error under $\boldsymbol{\mathcal{H}_0}$*
2. *Its power approaches $\mathbf{1}$ under $\boldsymbol{\mathcal{H}_1}$*

- The number of sieves shall grow to infinity $\longrightarrow$ reduce the model misspecification error (classical weak convergence theorem is **not** directly applicable)
- Develop a **matrix concentration inequality** under nonstationarity (sharper than naively applying concentration inequalities for scalar random variables)
- **Undersmoothing** is not needed to guarantee the test has good **size** property
- **Cross-validation** can be employed to select the number of sieves
- ℓ_1 and normalized maximum type tests require **weaker** conditions than the maximum-type test

Thank You!

☺Papers and softwares can be found on my personal website

`callmespring.githuo.io`