Proceedings of Machine Learning Research vol XX:1–16, 2023 5th Annual Conference on Learning for Dynamics and Control

Towards Dynamic Causal Discovery with Rare Events: A Nonparametric Conditional Independence Test

Chih-Yuan Chiu CHIHYUAN_CHIU@BERKELEY.EDU
337 Cory Hall, Berkeley CA 94709

Kshitij Kulkarni KSHITIJKULKARNI@BERKELEY.EDU
337 Cory Hall, Berkeley CA 94709

Shankar Sastry SASTRY@COE.BERKELEY.EDU
333B Cory Hall, Berkeley CA 94709

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

arXiv:2211.16596v2 [stat.ML] 1 Dec 2022

Abstract

Causal phenomena associated with rare events occur across a wide range of engineering problems, such as risk-sensitive safety analysis, accident analysis and prevention, and extreme value theory. However, current methods for causal discovery are often unable to uncover causal links, between random variables in a dynamic setting, that manifest only when the variables first experience low-probability realizations. To address this issue, we introduce a novel statistical independence test on data collected from time-invariant dynamical systems in which rare but consequential events occur. In particular, we exploit the time-invariance of the underlying data to construct a superimposed dataset of the system state before rare events happen at different timesteps. We then design a conditional independence test on the reorganized data. We provide non-asymptotic sample complexity bounds for the consistency of our method, and validate its performance across various simulated and real-world datasets, including incident data collected from the Caltrans Performance Measurement System (PeMS). Code containing the datasets and experiments is publicly available here.

Keywords: Causal Discovery, Time-Series Data, Rare Events, Conditional Independence Tests, Sample Complexity Bounds.

1. Introduction

The occurrence of rare yet consequential events during the evolution of a dynamical system is ubiquitous in many fields of engineering and science. Examples include natural disasters, vehicular accidents, and stock market crashes. When studying such phenomena, understanding the causal links between the disruptive event and the underlying system dynamics is crucial for controlling the system. In particular, if certain values of the system state increase the probability that the disruptive event occurs, control strategies should be implemented to steer the state away from such values. This can be accomplished, for instance, by incorporating a description of this causal relationship into the cost function that generates these control inputs in an optimization-based control framework. In general, it is important to consider the following question:

Main Question (Q) Given a rare event associated with the evolution of a dynamical system, does the onset of the event become more likely when the system state assumes certain values?

Below, we present a running example, invoked throughout ensuing sections to provide context.

© 2023 C.-Y. Chiu, K. Kulkarni & S. Sastry.

Running Example Consider the task of reducing the number of vehicular accidents on a road by identifying their causes. In particular, consider the scenario in which the amount of traffic on a network of roads has a causal effect on the occurrence of accidents. For example, on some busy streets, high traffic flow induces congestion and renders chain collisions more likely. In this case, since steady-state flows in a traffic network can be controlled via tolling, regulators can adjust the toll on each network link to redistribute flow and reduce the number of accidents that transpire Maheshwari et al. (2022a,b). Conversely, on other roads, low traffic flow may incentivize drivers to exceed the speed limit and create more opportunities for accidents to occur. In this case, traffic engineers can enforce speed limits more stringently at times of low traffic flow. ■

Although many well-established methods in the causal discovery literature can efficiently learn causal relationships from data, most only apply to data generated from probability distributions associated with static, acyclic Bayesian networks Glymour et al. (2019); Pearl (2009); Spirtes et al. (2000). Moreover, most causal discovery algorithms developed for time series data rely on stringent assumptions, such as linear dynamics and additive Gaussian noise models, or aggregate data along slices of fixed time indices Glymour et al. (2019); Gnecco et al. (2021); Granger (1969); Pérez-Ariza et al. (2012). However, rare events often occur sparsely at any fixed time and cannot be easily modeled using linear dynamics.

To address these shortcomings, we present a novel approach for aggregating and analyzing time series data in which consequential events of interest occur sparsely. Our method rests on the observation that, whereas a rare event may be highly unlikely to occur *at any fixed time* t, the probability of the event occurring *at some time along the entire horizon* of interest is often much higher. Thus, we aggregate the time series data along the times of the event's first occurrence. This renders the dataset more informative, by better representing the rare events of interest. Next, we present an algorithm that uses the curated data to analyze the causal relationships governing the occurrence of the rare event. The question of whether the system state affects the probability that the rare event occurs is formally posed as a binary hypothesis test, with the null hypothesis H_0 corresponding to the negative answer, and the alternative hypothesis H_1 corresponding to the positive one. We mathematically prove that our proposed method is *consistent against all alternatives* Lehmann (1951). In other words, if H_0 were true, then as the number of data trajectories N in the dataset approaches infinity, our approach would reject H_1 with probability 1. We validate the performance of our algorithm on simulated and on publicly available traffic and incident data collected from the Caltrans Performance Measurement System (PeMS).

2. Related Work

Causal Discovery for Static and Time Series Data Causal discovery algorithms identify causal links among a collection of random variables from a dataset of their realizations. Common approaches include constraint-based methods, which use statistical independence tests, score-based methods, which pose causal discovery as an optimization problem, and hybrid methods Glymour et al. (2019); Pearl (2009); Peters et al. (2017). However, most of these approaches apply only to non-temporal settings. For time series data, Granger causality uses vector autoregression to study whether one time series can be used to predict another Granger (1969). Other methods aggregate different data trajectories by matching time indices Pérez-Ariza et al. (2012); Entner and Hoyer (2010), or directly solve a time-varying causal graph Malinsky and Spirtes (2018). However, these methods do not address the problem of inferring causal links between rare events and dynamical systems, across sample trajectories on which the rare event can often occur at different times.

Extreme Value Theory and Analysis of Rare Events Extreme value theory characterizes dependences between random variables that exist only when a low-probability event occurs, e.g., rare meteorological events, or financial crises Engelke and Volgushev (2022); Asadi et al. (2018). Most closely related to our work are Gnecco et al. (2021), which studies causal links between heavy-tailed random variables, and Jana et al. (2021), which explores causal relationships between characteristics of London bicycle lanes, such as density, length, and collision rate, and abnormal congestion. However, Gnecco et al. (2021) imposes restrictive assumptions, such as linear models, while the discussion in Jana et al. (2021) on accidents' occurrences is restricted to empirical studies. In contrast, our proposed algorithm applies a nonparametric conditional independence test that is capable of inferring relationships between a general dynamical system, and the onset of a rare event.

Traffic Network Analysis Traffic network theory aims to mathematically describe and control traffic flow in urban networks of roads, bridges, and highways Baillon and Cominetti (2008); Krichene et al. (2014); Ahipaşaoğlu et al. (2019). Recent literature has proposed the design of tolling mechanisms that drive a traffic network to the socially optimal steady state Maheshwari et al. (2022b); Como and Maggistro (2022). However, these methods do not model or predict the occurrence of sudden yet consequential events, such as extreme weather events, car accidents, and other causes of unexpected congestion. In contrast, our paper uses the occurrence of rare but consequential car accidents in traffic networks as a running example, to illustrate the applicability of our method on analyzing causal links between dynamical systems and associated rare events.

3. Preliminaries

Consider a stochastic, discrete-time dynamical system with state variable $X_t \in \mathbb{R}^n$, event variable $A_t \in \{0,1\}$ with $\mathbb{P}(A_t = 1) \in [p_1, p_2]$ for some $p_1, p_2 \in (0,1)$ for all t, with $p_1 < p_2$, and dynamics $X_{t+1} = f(X_t, A_t, W_t)$ for each $t \geq 0$, where $W_t \in \mathbb{R}^w$ denotes i.i.d. noise, and $f : \mathbb{R}^n \times \{0,1\} \times \mathbb{R}^w \to \mathbb{R}^n$ denotes the nonlinear dynamics of the system state. Let T denote the time at which the rare event first occurs, and, with a slight abuse of notation,let $A_{1:t} = 0$ denote the event that $A_1 = \cdots = A_t = 0$. Moreover, we assume that the first occurrence of the rare event is governed by a time-invariant probability distribution, i.e.,:

$$\mathbb{P}(A_{t+1} = 1 | X_t \preceq x, A_{1:t} = 0) = \mathbb{P}(A_{t'+1} = 1 | X_{t'} \preceq x, A_{1:t'} = 0), \qquad \forall t, t' \geq 0, \quad (1)$$

where, for each $x, y \in \mathbb{R}^n$, the notation $x \preceq y$ represents $x_i \leq y_i$ for each $i \in [n] := \{1, \cdots, n\}$, and for each $x \in \mathbb{R}^n$, there exists some constant ratio $\alpha(x) > 0$ such that $\mathbb{P}(X_{t-1} \preceq x | A_t = 1, A_{1:t-1} = 0) = \alpha(x) \cdot \mathbb{P}(X_{t-1} \preceq x | A_{1:t-1} = 0)$. Moreover, we assume that for each $x \in \mathbb{R}^n$, there exists some $\alpha(x) \geq 0$ such that:

$$\mathbb{P}(X_{t-1} \preceq x | A_t = 1, A_{1:t-1} = 0) = \alpha(x) \cdot \mathbb{P}(X_{t-1} \preceq x | A_{1:t-1} = 0). \quad (2)$$

In words, we assume that the flow distribution is related to the first occurrence of the rare event in a time-invariant manner. Given this setup, we restate **Q**, first defined in the introduction, as the following hypothesis testing problem:

Definition 1 *The binary hypothesis test, with null hypothesis H_0 defined below, is a mathematically rigorous characterization of **Q**:*

$$H_0 : \qquad \mathbb{P}(A_{t+1} = 1 | X_t \preceq x, A_{1:t} = 0) = \mathbb{P}(A_{t+1} = 1 | A_{1:t} = 0), \qquad \forall x \in \mathbb{R}.$$

In words, H_0 holds if and only if the first occurrence of the rare event transpires independently of the system state at that time. For convenience, we define the left and right hand sides of H_0 by:

$$a_1(x) := \mathbb{P}(A_{t+1} = 1 | X_t \preceq x, A_{1:t} = 0), \tag{3}$$

$$a_2 := \mathbb{P}(A_{t+1} = 1 | A_{1:t} = 0). \tag{4}$$

Running Example Consider a parallel link traffic network of R links that connect a single source and a single destination. Let $X_{t,i} \in \mathbb{R}$ denote the traffic flow on every link $i \in [R] := \{1, \cdots, R\}$ at time t, and define $X_t := (X_{t,1}, \cdots, X_{t,r}) \in \mathbb{R}^R$. (In general, one can define $X_{t,i} \in \mathbb{R}^d$ to encapsulate other observed quantities relevant to link i at time t, e.g., vehicle speed and pavement quality). The event variable $A_t = 1$ corresponds to the occurrence of an accident in the network at time t.

In this context, Definition 1 corresponds to checking whether the first occurrence of an accident on the R-link network at time t is affected by the flow level at time $t-1$. This question may be of interest to traffic authorities, since if the occurrences of costly accidents becomes more likely at certain levels of traffic flow X_t, then the flow should be monitored to decrease the chance that such accidents occur. Flow management can be applied by dynamically tolling the links, as in Maheshwari et al. (2022a). As accidents are relatively rare in most traffic datasets, it can be difficult to construct accurate estimates of accident probabilities and flows before accidents at any given time t. Instead, below, we propose a novel method of data aggregation that allows the use of information on accident occurrences across all times. ■

Since X_t is a continuous random variable, a direct comparison of (3) and (4) would necessitate computing (3) for uncountably many values of $x \in \mathbb{R}^n$. Instead, we use the laws of conditional and total probability to reformulate the problem. In the spirit of Bayes' rule, we compare the state distribution immediately before the rare event occurred, instead of the rare event probabilities under different state values. Formally, we observe that under either hypothesis, the state distribution immediately before the first accident can be decomposed as the following infinite sum; for each $x \in \mathbb{R}^n$:

$$\begin{aligned}
\mathbb{P}(X_{T-1} \preceq x) &= \sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, T = t) \\
&= \sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, A_t = 1, A_{1:t-1} = 0) \\
&= \sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, A_{1:t-1} = 0) \cdot \mathbb{P}(A_t = 1 | X_{t-1} \preceq x, A_{1:t-1} = 0).
\end{aligned}$$

Intuitively, if H_0 were true, then the condition $X_{t-1} \preceq x$ in the term $\mathbb{P}(A_t = 1 | X_{t-1} \preceq x, A_{1:t-1} = 0)$ can be dropped. This observation is rigorously formulated as Proposition 2, stated below.

Proposition 2 *The null hypothesis H_0 in Definition 1 holds if and only if, for each $x \in \mathbb{R}^n$:*

$$\mathbb{P}(X_{T-1} \preceq x) = \sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, A_{1:t-1} = 0) \cdot \mathbb{P}(A_t = 1 | A_{1:t-1} = 0). \tag{5}$$

Proof Please see Appendix A in the ArXiV version of the paper (link). ■

For convenience, we define, for each $t \in \mathbb{N}$ and $x \in \mathbb{R}$:

$$\begin{aligned} b_1(x) &:= \mathbb{P}(X_{T-1} \preceq x), \\ \beta_t(x) &:= \mathbb{P}(X_{t-1} \preceq x, A_{1:t-1} = 0), \\ \gamma_t &:= \mathbb{P}(A_t = 1 | A_{1:t-1} = 0), \\ b_2(x) &:= \sum_{t=1}^{\infty} \beta_t(x) \cdot \gamma_t = \sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, A_{1:t-1} = 0) \cdot \mathbb{P}(A_t = 1 | A_{1:t-1} = 0) \\ &= \sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x | A_{1:t-1} = 0) \cdot \mathbb{P}(T = t). \end{aligned}$$

The test statistic that we use to distinguish between the distributions $b_1(x)$ and $b_2(x)$ is the maximum CDF gap between the distributions, i.e.,:

$$\sup_{x \in \mathbb{R}^n} |b_1(x) - b_2(x)|$$

Intuitively, a large gap indicates a higher likelihood that a component-wise larger or smaller state would change the probability of an event occurring. We formalize this notion in Algorithm 1, and provide finite sample guarantees for empirical estimates of $b_1(x)$ and $b_2(x)$ that can be constructed efficiently from data and used to compute the test statistic.

Running Example In the traffic network example, $b_1(x)$ corresponds to the probability that X_{T-1}, the network flows before the first accident, is component-wise less than or equal to x. Meanwhile, $b_2(x)$ describes the weighted average of traffic flows at each time t, conditioned on the first accident occurring after t, with the distribution of the first accident time T as weights. Section 4 describes sample-efficient methods for constructing empirical estimates of $b_1(x)$ and $b_2(x)$ from a dataset of independent traffic flows. ■

4. Methods

4.1. Main Algorithm

We present Algorithm 1, which solves the hypothesis testing problem in Definition 1 from a dataset of N independent trajectories, by constructing and comparing finite-sample empirical cumulative distribution functions (CDFs) $\hat{b}_1^N(x)$ and $\hat{b}_2^N(x)$ for the expressions $b_1(x)$ and $b_2(x)$, respectively, and verifying whether or not (5) holds (in accordance with Proposition 2).

Algorithm 1: Hypothesis Testing with Reorganized Dataset

Data: Dataset of system state and rare event variables: $\{(X_t^i, A_t^i) : t \geq 0, i \in [N]\}$
Result: Distribution gap: $\sup_{x\in\mathbb{R}} |\hat{b}_1^N(x) - \hat{b}_2^N(x)|$

$\hat{T}^i \leftarrow$ Realization of T for data trajectory i, $\forall\, i \in [N]$

$\hat{b}_1^N(x) \leftarrow \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{X_{\hat{T}^i} \preceq x\}$.

$\hat{\beta}_t^N(x) \leftarrow \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{X_t^i \preceq x, A_{1:t-1}^i = 0\}$.

$$\hat{\gamma}_t^N \leftarrow \begin{cases} \frac{\sum_{i=1}^N \mathbf{1}\{A_{1:t}^i = 0\}}{\sum_{i=1}^N \mathbf{1}\{A_{1:t-1}^i = 0\}}, & \text{if } \sum_{i=1}^N \mathbf{1}\{A_{1:t-1}^i = 0\} > 0, \\ 0, & \text{else.} \end{cases}.$$

$\hat{b}_2^N(x) \leftarrow \sum_{t=1}^\infty \hat{\beta}_t^N(x) \cdot \hat{\gamma}_t^N$.

Return $\sup_{x\in\mathbb{R}} |\hat{b}_1^N(x) - \hat{b}_2^N(x)|$.

Note on the baseline method The common baseline method for resolving the problem in Definition 1 is to fix $t \geq 1$, and compare the CDF values $\mathbb{P}(X_{t-1} \preceq x | T = t)$ and $\mathbb{P}(X_{t-1} \preceq x)$, for each $x \in \mathbb{R}^n$ at the fixed t. This is effectively a "static variant" of Algorithm 1 that only utilizes dynamical state values immediately before accidents that occur at time t. It is generally difficult to estimate $\mathbb{P}(X_{t-1} \preceq x | T = t)$ from data, since $\mathbb{P}(T = t)$ can be very small for any given t. Our algorithm (Algorithm 1) instead aggregates data *across* times when the rare event has occurred, allowing the event to be better represented.

4.2. Theoretical Guarantees:

Theorem 3 below illustrates that, if H_0 holds, then as the number of sample trajectories N approaches infinity, the empirical distributions of (6) and (7), as constructed in Algorithm 1 converge at an exponential rate to their true values. This establishes a finite sample bound that controls the error of the statistical independence test presented in Algorithm 1. The proof follows by carefully applying concentration bounds for light-tailed random variables, and invoking the Dvoretsky-Kiefer-Wolfowitz inequality, which prescribes explicit convergence rates for empirical CDFs.

Theorem 3 (Exponential Convergence to Consistency Against all Alternatives) *Suppose the null hypothesis H_0 holds, i.e., $b_1(x) = b_2(x)$.*

1. *If $n = 1$, i.e., $X_t \in \mathbb{R}$ for each $t \geq 0$, then for each $\epsilon > 0$, there exist continuous, positive functions $C_1(\epsilon), C_2(\epsilon) > 0$ such that:*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left\{\left|\hat{b}_1^N(x) - \hat{b}_2^N(x)\right|\right\} > \epsilon\right) \leq C_1(\epsilon) \cdot e^{-N\cdot C_2(\epsilon)}.$$

2. *If $n > 1$, then there exist continuous, positive functions $C_3(\epsilon), C_4(\epsilon) > 0$ such that:*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}^n} \left\{\left|\hat{b}_1^N(x) - \hat{b}_2^N(x)\right|\right\} > \epsilon\right) \leq \left[C_3(\epsilon)(N+1)n + C_4(\epsilon)\right] \cdot e^{-N\cdot C_5(\epsilon)}.$$

Moreover, for sufficiently large N, the factor $N+1$ can be replaced by the constant 2.

Proof Please see Appendix B in the ArXiV version of the paper (link). ■

5. Results

In this section, we illustrate the numerical performance of our proposed method on simulated and real-world traffic data, and its efficacy over baseline aggregation methods of concatenating data points along a single, fixed time t.

5.1. Simulated Data

In our first set of experiments, we construct synthetic data for single-link and multi-link traffic networks. For the single-link network, we use the dynamics:

$$\begin{aligned} x[t+1] &= (1-\mu(A[t]))\cdot x[t] + \mu(A[t])\cdot u[t] + w[t], \qquad \forall\, t \in [T_h], \\ A[t+1] &\sim \mathcal{P}(x[t]) \end{aligned}$$

where $x[t] \in \mathbb{R}$ denotes the traffic flow at time t, $A[t] \in \{0,1\}$ is the Boolean random variable that indicates whether or not an accident has occurred at time t, $\mu(A[t]) > 0$ describes the fraction of traffic flow departing the link, $u[t] \in \mathbb{R}$ denotes the total input traffic flow, $w[t] \in \mathbb{R}$ is a zero-mean noise term, and T_h is the finite time horizon. In our experiments, we set $T_h = 500$, $\mu(0) = 0.3$, $\mu(1) = 0.2$, $u(t) = 100$ for each $t \in [T_h]$, and draw $w(t)$ i.i.d. from the continuous uniform distribution on $(-10, 10)$. We create datasets corresponding to the null and alternative hypothesis. For the null hypothesis dataset, we fix $\mathcal{P}(x[t]) =$ Bernoulli(0.01), regardless of the value of $x[t]$. This simulates a scenario where the likelihood of an accident occurring has no dependence on traffic flow. For the alternative hypothesis dataset, we set $\mathcal{P}(x[t]) =$ Bernoulli(0.01) when $x[t] < 109$ and $\mathcal{P}(x[t]) =$ Bernoulli(0.10) when $x[t] \geq 109$. This represents a scenario where higher traffic loads increase the likelihood that an accident occurs.

To contrast the performance of our algorithm with the baseline, we compute the following quantities from datasets of independent trajectories corresponding to H_0 and H_1, in accordance with Proposition 2 and Theorem 3:

- **For our method**—We compute the empirical estimates $\hat{b}_1^N(x)$ and $\hat{b}_2^N(x)$ of the functions $b_1(x)$ and $b_2(x)$ as functions of x (Figure 1), and the maximum CDF gap $\sup_{x\in\mathbb{R}} |\hat{b}_1^N(x) - \hat{b}_2^N(x)|$ as functions of N (Figure 2).
- **For the baseline method**—We compute the empirical estimates of the CDFs of $X_{t-1}|T = t$ and X_{t-1}, with t fixed at 1, as functions of x (Figure 1), and the corresponding maximum CDF gap as functions of N (Figure 2). Note that for $N < 500$, it is difficult to obtain the CDF of $X_{t-1}|T = t$, due to rarity of the event at any given time.

Figures 1 and 2 demonstrate that, compared to the baseline method, our approach is able to distinguish between the null and alternative hypotheses from a far smaller dataset. This illustrates that our method, compared to the baseline, distinguishes the dependence between the occurrence of a rare event and the state values immediately preceding the event more efficiently.

Appendix C contains further empirical results on synthetic datasets for multi-link networks.

5.2. Caltrans PeMS Dataset

We now demonstrate the efficacy of our algorithm by conducting the hypothesis test in Proposition 2 on real traffic flow and incident data collected from the publicly available Caltrans Performance

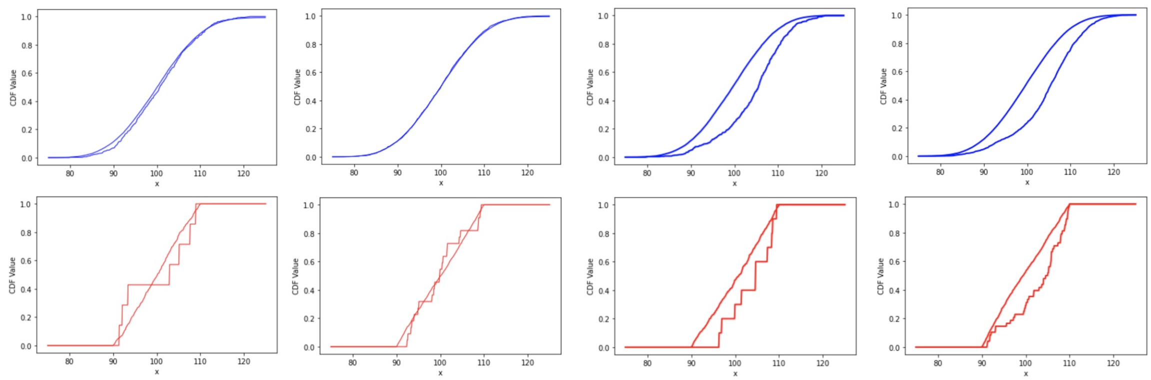


Figure 1: (Top) From left to right, $b_1(x)$ and $b_2(x)$ vs. x plots for $(H_0, N = 500)$, $(H_0, N = 2000)$, $(H_1, N = 500)$, and $(H_1, N = 2000)$. (Bottom) From left to right, empirical CDFs for $X_{t-1}|T = t$ and X_{t-1} with $t = 1$, in the same order of hypothesis and N values.

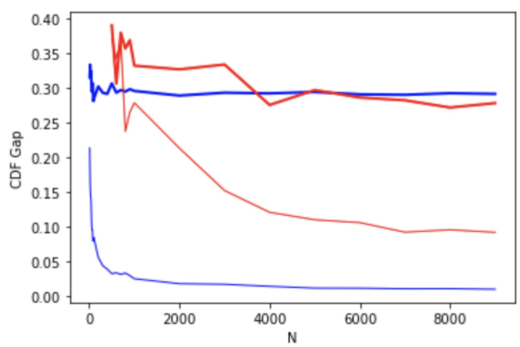


Figure 2: CDF gap between vs. N. Here, red and blue correspond to the baseline and our method, respectively, while thick and thin lines correspond to the null and alternative hypotheses, respectively. Our approach (thin blue curve) correctly identifies the null hypothesis dataset with a relatively small number of samples, while the baseline aggregation method fails to do so (thin red curve).

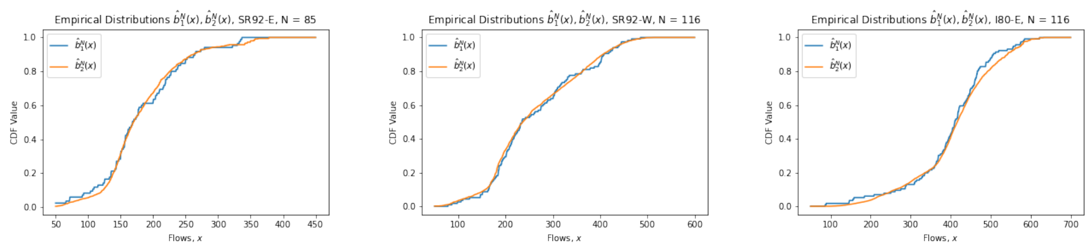
Measurement System (PeMS) dataset Varaiya (2001). PeMS uses loop detectors placed on freeways to collect flow, speed, and other data about traffic conditions, and overlays this with incident reports. We consider traffic flow data (# vehicles / time) collected from January to August 2022 from 6 A.M. to 2 P.M. daily, at 5-minute intervals, on a selection of bridges: the San Mateo-Hayward Bridge, the San Francisco–Oakland Bay Bridge, and the Richmond-San Rafael Bridge, in the San Francisco Bay Area. That is, we consider single link networks connecting a source and destination with the continuous variables $X_t \in \mathbb{R}_+$ corresponding to average flows on the link. Correspondingly, we use incident data collected on these bridges by PeMS in the same time interval from the California Highway Patrol (CHP).

Data Collection We treat each day as an independent trajectory of the traffic flows on every bridge. The PeMS dataset contains flows collected from dual loop detectors placed along the bridges. For each time between 6 A.M. and 2 P.M., we average the flow data recorded by loop detectors on each bridge to obtain the state variable X_t for time t. Mathematically, we define $X_t := \frac{1}{|I|} \sum_{i=1}^{|I|} X_t^i$, where I denotes the set of loop detectors on a single link, and X_t^i denotes the flow measured by detector $i \in I$ at time t. We exclude from our analysis any trajectory on which there was no incident for the entire day, since such trajectories do not contain data relevant to our problem of interest.

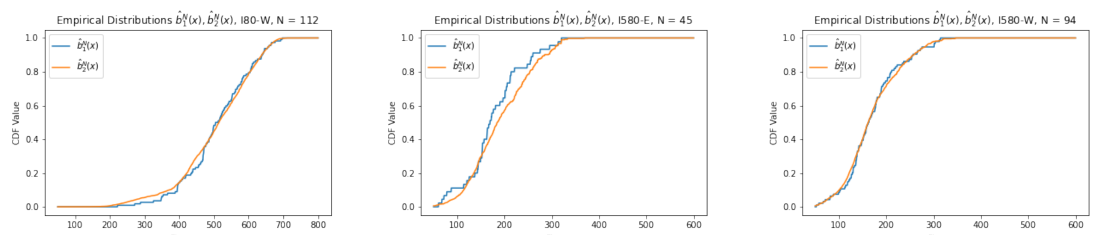
Results In Table 1, we enumerate the sample size N and test statistic $\sup_{x \in \mathbb{R}} |\hat{b}_1^N(x) - \hat{b}_2^N(x)|$ for the six traffic links (three bridges, each with two directions of traffic flow). Note the substantial difference in the CDF gap (of nearly 0.178) for the Richmond-San Rafael Bridge, East, compared to all other links, indicating that the flows on this link are particularly causally linked to the first time of incident formation. Further, the San Francisco-Oakland Bay Bridge, East, also has a higher CDF gap (0.081) relative to the West direction, and relative to the other bridges. These gaps are visible in the CDF plots in Figures 3(e) and 3(c), respectively.

Link	N	$\sup_{x \in \mathbb{R}} \|\hat{b}_1^N(x) - \hat{b}_2^N(x)\|$
San Mateo-Hayward Bridge, East (SR92-E)	85	0.053
San Mateo-Hayward Bridge, West (SR92-W)	116	0.039
San Francisco–Oakland Bay Bridge, East (I80-E)	116	0.081
San Francisco–Oakland Bay Bridge, West (I80-W)	112	0.042
Richmond-San Rafael Bridge, East (I580-E)	45	0.178
Richmond-San Rafael Bridge, West (I580-W)	94	0.048

Table 1: CDF gap, $\sup_{x \in \mathbb{R}} |\hat{b}_1^N(x) - \hat{b}_2^N(x)|$ for the six links in the San Francisco Bay Area.



(*a*) Sans Mateo-Hayward Bridge, East (SR92-E) (*b*) San Mateo-Hayward Bridge, West (SR92-W) (*c*) San Francisco–Oakland Bay Bridge, East (I80-E)



(*d*) San Francisco–Oakland Bay Bridge, West (I80-W) (*e*) Richmond-San Rafael Bridge, East (I580-E) (*f*) Richmond-San Rafael Bridge, West (I580-W)

Figure 3: Empirical CDFs $\hat{b}_1^N(x)$ and $\hat{b}_2^N(x)$ for six bridges in the San Francisco Bay Area.

6. Conclusion

We present a novel method for identifying causal links between the state evolution of a dynamical system and the onset of an associated rare event. Crucially, we leverage the time-invariance of frequently encountered dynamical models to reorganize data in a manner that better represents occurences of the rare event. We then formulate a nonparametric statistical independence test to infer causal dependencies between the states of the dynamical system and the rare event. Empirical results on simulated and real-world time-series data indicate that our method outperforms a baseline approach that conducts independence tests only a single time slice of the original dataset, in which rare events occur sparsely. Further, when applied to real-world traffic flow and incident data collected from the Caltrans PeMS system, our method indicates that there are indeed bridges in the San Francisco Bay Area on which accident occurrences are causally linked to traffic flow.

As future work, the causal discovery algorithm presented here may be used to more effectively control the evolution of a dynamical system associated with a rare but consequential event. By establishing causal links between the dynamical state and the rare event, control strategies can be redesigned to maneuver the state away from regions of the state space where the event occurs more frequently. Important engineering applications include incentive design and flow control methods in the network traffic systems literature, such as dynamic tolling and rerouting. In addition, it is of interest to develop more flexible formulations of our method to study more complicated models of rare events, such as continuous and multivariate variables that account for the severity of the event, including the severity of accidents on a traffic network.

The appendix can be found on ArXiV (link), per L4DC submission instructions. The authors will ensure that the ArXiV link stays active.

References

Selin Damla Ahipaşaoğlu, Uğur Arıkan, and Karthik Natarajan. Distributionally Robust Markovian Traffic Equilibrium. *Transportation Science*, 53(6):1546–1562, 2019.

Peiman Asadi, Sebastian Engelke, and Anthony C. Davison. Optimal Regionalization of Extreme Value Distributions for Flood Estimation. *Journal of Hydrology*, 556:182–193, 2018.

Jean-Bernard Baillon and Roberto Cominetti. Markovian Traffic Equilibrium. *Mathematical Programming*, 02 2008. doi: 10.1007/s10107-006-0076-2.

Giacomo Como and Rosario Maggistro. Distributed Dynamic Pricing of Multiscale Transportation Networks. *IEEE Transactions on Automatic Control*, 67(4):1625–1638, 2022.

Sebastian Engelke and Stanislav Volgushev. Structure Learning for Extremal Tree Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2022.

Doris Entner and Patrik O Hoyer. On Causal Discovery from Times Series Data Using FCI. *Probabilistic graphical models*, pages 121–128, 2010.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10, 2019. doi: 10.3389/fgene.2019.00524.

Nicola Gnecco, Nicolai Meinshausen, Jonas Peters, and Sebastian Engelke. Causal Discovery in Heavy-tailed Models. *The Annals of Statistics*, 49(3):1755–1778, 2021.

Clive William John Granger. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. In *Econometrica*, volume 37, pages 424–438, 1969.

Kaushik Jana, Prajamitra Bhuyan, and Emma McCoy. Causal Analysis at Extreme Quantiles with Application to London Traffic Flow Data. *ArXiv*, 08 2021.

Walid Krichene, Benjamin Drighès, and Alexandre Bayen. On the Convergence of No-Regret Learning in Selfish Routing. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 163–171, Bejing, China, 22–24 Jun 2014. PMLR.

Erich Leo Lehmann. Consistency and Unbiasedness of Certain Nonparametric Tests. *Annals of Mathematical Statistics*, 22:165–179, 1951.

Chinmay Maheshwari, Kshitij Kulkarni, Manxi Wu, and S. Shankar Sastry. Dynamic Tolling for Inducing Socially Optimal Traffic Loads. In *2022 American Control Conference (ACC)*, pages 4601–4607, 2022a. doi: 10.23919/ACC53348.2022.9867193.

Chinmay Maheshwari, Kshitij Kulkarni, Manxi Wu, and S. Shankar Sastry. Inducing Social Optimality in Games via Adaptive Incentive Design. *ArXiv*, abs/2204.05507, 2022b.

Daniel Malinsky and Peter L. Spirtes. Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding. In *CD@KDD*, 2018.

Michael Naaman. On the Tight Constant in the Multivariate Dvoretzky–Kiefer–Wolfowitz Inequality. *Statistics & Probability Letters*, 173:109088, 2021. ISSN 0167-7152.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge university press, 2009.

Cora Beatriz Pérez-Ariza, Ann E. Nicholson, Kevin B. Korb, Steven Mascaro, and Chao Heng Hu. Causal Discovery of Dynamic Bayesian Networks. In *Proceedings of the 25th Australasian Joint Conference on Advances in Artificial Intelligence*, AI'12, page 902–913, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 9783642351006. doi: 10.1007/978-3-642-35101-3_76.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. ISBN 978-0-262-03731-0.

P. Spirtes, Clark Glymour, Richard Scheines, Stuart A. Kauffman, Valerio Aimale, and Francis C Wimberly. Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data. In *Proceedings of the Atlantic Symposium on Computational Biology*, 2000.

Pravin Varaiya. Freeway performance measurement system: Final report. *Univ. of California Berkeley, Berkeley, CA, USA, Tech. Rep. UCB-ITSPWP-2001-1*, 2001.

Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018. ISBN 1108415199.

Appendix A. Preliminaries

Proof (**Proof of Proposition 2**) First, for each $x \in \mathbb{R}^n$:

$$\begin{aligned}
&\mathbb{P}(A_t = 1 | X_{t-1} \leq x, A_{1:t-1} = 0) \\
=&\mathbb{P}(A_t = 1 | A_{1:t-1} = 0) \cdot \frac{\mathbb{P}(X_{t-1} \preceq x | A_t = 1, A_{1:t-1} = 0)}{\mathbb{P}(X_{t-1} \preceq x | A_{1:t-1} = 0)} \\
=&\mathbb{P}(A_t = 1 | A_{1:t-1} = 0) \cdot \alpha(x).
\end{aligned}$$

Thus, $\mathbb{P}(A_t = 1 | A_{t-1} = 0)$ is time-invariant. Next, observe that:

$$\begin{aligned}
&\mathbb{P}(X_{T-1} \preceq x) \\
=&\sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x | T = t) \cdot \mathbb{P}(T = t) \\
=&\sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, T = t) \\
=&\sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, A_{1:t-1} = 0, A_t = 1) \\
=&\sum_{t=1}^{\infty} \mathbb{P}(A_t = 1 | X_t \preceq x, A_{1:t-1} = 0) \cdot \mathbb{P}(X_{t-1} \preceq x, A_{1:t-1} = 0) \\
=&a_1(x) \cdot \sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, A_{1:t-1} = 0).
\end{aligned} \tag{6}$$

and:

$$\begin{aligned}
&\sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, A_{1:t-1} = 0) \cdot \mathbb{P}(A_t = 1 | A_{1:t-1} = 0) \\
=&a_2 \cdot \sum_{t=1}^{\infty} \mathbb{P}(X_{t-1} \preceq x, A_{1:t-1} = 0).
\end{aligned} \tag{7}$$

Thus, the null hypothesis H_0 in Definition 1 holds if and only if (6) and (7) are equal, as claimed. ■

Appendix B. Methods

Proposition 4 (Corollary to the Dvoretzky-Kiefer-Wolfowitz Inequality) *Let $X \in \mathbb{R}^n$ be a random variable defined on the probability space $(\Omega, \Sigma, \mathbb{P})$, and let $E \in \Sigma$. Fix $\epsilon > 0$ and $N \in \mathbb{N}$, and let $X_1, \cdots, X_N$ be i.i.d. copies of X. Then, for each $N \geq \mathbb{N}$ and each $\epsilon > 0$:*

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}^n} \left| \mathbb{P}(X \preceq x, E) - \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{X_i \preceq x, E\} \right| \right) \leq 2e^{-2N\epsilon^2}.$$

Proof Let $G_1, \cdots, G_N$ be drawn i.i.d. from the continuous uniform $(0,1)$ distribution. Then:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathbb{R}^n}\left|\mathbb{P}(X\preceq x, E) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{X_i\preceq x, E\}\right| > \epsilon\right) \\
=&\mathbb{P}\left(\sup_{x\in\mathbb{R}^n}\left|\mathbb{P}(X\preceq x, E) - G_n\big(\mathbb{P}(X\preceq x, E)\big)\right| > \epsilon\right) \\
\leq&\mathbb{P}\left(\sup_{t\in[0,1]}\left|t - G_n(t)\right| > \epsilon\right) \\
\leq& 2e^{-2N\epsilon^2},
\end{aligned}$$

where the final inequality follows by applying the Dvoretzky-Kiefer-Wolfowitz Inequality to the continuous uniform $(0,1)$ distribution. ■

Proof Fix $\epsilon > 0$, and take:

$$T_c(\epsilon) := \left\lceil \frac{1}{\ln(1-p_1)}\ln\left(\frac{\epsilon p_1^2}{16 p_2}\right)\right\rceil.$$

First, to show that $\hat{b}_1^N(x) \to b_1(x)$ at an exponential rate in N, we invoke the Dvoretzky-Kiefer-Wolfowitz inequality:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\hat{b}_1^N(x) - b_1(x)\right| > \frac{1}{2}\epsilon\right) \leq 2\cdot e^{-\frac{1}{2}N\epsilon^2}$$

Next, to show that $\hat{b}_2^N(x) \to b_2(x)$ at an exponential rate in N, we have:

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\left\{\hat{b}_2^N(x) - b_2(x)\right\}\right| \\
=&\sup_{x\in\mathbb{R}}\left|\sum_{t=1}^{\infty}\left[\hat{\beta}_t^N(x)\hat{\gamma}_t^N - \beta_t(x)\gamma_t\right]\right| \\
=&\sum_{t=1}^{T_c}\left[\sup_{x\in\mathbb{R}}\left\{|\hat{\beta}_t^N(x) - \beta_t(x)|\right\}\hat{\gamma}_t^N + \sup_{x\in\mathbb{R}}\left\{|\hat{\gamma}_t^N - \gamma_t|\right\}\beta_t(x)\right] \\
&\quad + \sup_{x\in\mathbb{R}}\left\{\sum_{t=T_c+1}^{\infty}\left[|\hat{\beta}_t^N(x)\hat{\gamma}_t^N| + |\beta_t(x)\gamma_t|\right]\right\} \\
\leq&\sum_{t=1}^{T_c}\sup_{x\in\mathbb{R}}\left\{|\hat{\beta}_t^N(x) - \beta_t(x)|\right\} + \sum_{t=1}^{T_c}\sup_{x\in\mathbb{R}}\left\{|\hat{\gamma}_t^N - \gamma_t|\right\}\cdot\mathbb{P}(A_{1:t-1}=0) \\
&\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{t=T_c+1}^{\infty}\mathbf{1}\{\hat{T}^n \geq t\} + \sum_{t=T_c+1}^{\infty}\mathbb{P}(T\geq t).
\end{aligned}$$

Below, we upper bound each of the four terms in the final expression above.

- First, by the Dvoretzky-Kiefer-Wolfowitz inequality, we have, for each $t \in [T_c] := \{1, \cdots, T_c\}$:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{t=1}^{T_c} \sup_{x\in\mathbb{R}} \left\{|\hat{\beta}_t^N(x) - \beta_t(x)|\right\} \geq \frac{1}{8}\epsilon\right) \\
\leq &\sum_{t=1}^{T_c} \mathbb{P}\left(\sup_{x\in\mathbb{R}} \left\{|\hat{\beta}_t^N(x) - \beta_t(x)|\right\} \geq \frac{1}{8T_c}\epsilon\right) \\
\leq &\, 2T_c \exp\left(-\frac{\epsilon^2}{32T_c^2} \cdot N\right).
\end{aligned}$$

- Let $N_t \in [N]$ denote the number of trajectories with $A_{1:t-1} = 0$. We first show that, with high probability, $N_t \geq N \cdot \mathbb{P}(A_{1:t-1} = 0)^2$. We then show that, under this condition on N_t taking a sufficiently large value, $\hat{\gamma}_t^N(x) \to \gamma_t(x)$ exponentially in N.

  First, by the Hoeffding bound for general bounded random variables (Vershynin Vershynin (2018), Theorem 2.2.6), we have:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{N}N_t \leq \mathbb{P}(A_{1:t-1} = 0)^2\right) \\
\leq &\mathbb{P}\left(\left|\frac{1}{N}N_t - \mathbb{P}(A_{1:t-1} = 0)\right| > \mathbb{P}(A_{1:t-1} = 0) - \mathbb{P}(A_{1:t-1} = 0)^2\right) \\
\leq &\exp\left(-2\Big[\mathbb{P}(A_{1:t-1} = 0) - \mathbb{P}(A_{1:t-1} = 0)^2\Big]^2 \cdot N\right)
\end{aligned}$$

  Then, if $N_t \geq N \cdot \mathbb{P}(A_{1:t-1} = 0)$:

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\gamma}_t^N(x) - \gamma_t(x)| > \frac{\epsilon}{8T_c \cdot \mathbb{P}(A_{1:t-1} = 0)}\right) \\
\leq &\exp\left(-2 \cdot \mathbb{P}(A_{1:t-1} = 0)^2 \cdot N \cdot \frac{\epsilon^2}{64T_c^2 \cdot \mathbb{P}(A_{1:t-1} = 0)^2}\right) \\
\leq &\exp\left(-\frac{\epsilon^2}{32T_c^2} \cdot N\right).
\end{aligned}$$

- To bound the third term, $\frac{1}{N}\sum_{n=1}^{N}\sum_{t=T_c+1}^{\infty} \mathbf{1}\{\hat{T}^n \geq t\}$, define:

$$\begin{aligned}
B_{T_c} := \sum_{\tau=T_c+1}^{\infty} \mathbf{1}\{T \geq t\} &= \sum_{\tau=T_c+1}^{\infty}\sum_{\tau=t}^{\infty} \mathbf{1}\{T = \tau\} \\
&= \sum_{\tau=T_c+1}^{\infty}\sum_{\tau=t}^{\infty} \mathbf{1}\{T = \tau\} = \sum_{\tau=T_c+1}^{\infty}\sum_{t=T_c+1}^{\tau} \mathbf{1}\{T = \tau\} \\
&= \sum_{\tau=T_c+1}^{\infty} (\tau - T_c)\mathbf{1}\{T = \tau\} = \sum_{\tau=1}^{\infty} \tau \cdot \mathbf{1}\{T = \tau + T_c\}.
\end{aligned}$$

  Thus, we have:

$$\mathbb{E}[B_{T_c}] = \sum_{\tau=1}^{\infty} \tau \cdot \mathbb{P}(T = \tau + T_c) \leq \sum_{\tau=1}^{\infty} \tau \cdot (1 - p_1)^{\tau+T_c-1} p_2$$

$$\leq (1-p_1)^{T_c} \cdot \frac{p_2}{p_1} \cdot \sum_{\tau=1}^{\infty} \tau(1-p_1)^{\tau-1} p_1$$
$$= \frac{(1-p_1)^{T_c} \cdot p_2}{p_1^2},$$

and:

$$\mathbb{P}(B_{T_c} \leq t) \leq (1-p_1)^{t+k-1} p_2 \leq 2\exp\left(-\ln\left(\frac{1}{1-p_1}\right) t\right).$$

Thus, $B_{T_c} - \mathbb{E}[B_{T_c}]$ is sub-exponential, with: (see Vershynin Vershynin (2018), Proposition 2.7.1)

$$\|B_{T_c} - \mathbb{E}[B_{T_c}]\|_{\psi_1} = 6\sqrt{\frac{e}{\ln 2}}\left(1 + \frac{6}{\ln 2}\right) \cdot \frac{-1}{\ln(1-p_1)}.$$

Applying the Bernstein inequality for zero-mean sub-exponential variables (Vershynin, Vershynin (2018)), we obtain:

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} B_{T_c}^N > \frac{1}{8}\epsilon\right)$$
$$= \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} B_{T_c}^N - \mathbb{E}[B_{T_c}] > \frac{1}{8}\epsilon - \mathbb{E}[B_{T_c}]\right)$$
$$\leq \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} B_{T_c}^N - \mathbb{E}[B_{T_c}] > \frac{1}{16}\epsilon\right)$$
$$= \exp\Bigg(-\min\Bigg\{-\frac{1}{96}\sqrt{\frac{\ln 2}{e}}\left(\frac{\ln 2}{6+\ln 2}\right)\ln(1-p_1)C_1 \cdot \epsilon,$$
$$\frac{\ln 2}{9216e}\left(\frac{\ln 2}{6+\ln 2}\right)^2 \left[\ln(1-p_1)\right]^2 C_2 \cdot \epsilon^2\Bigg\} \cdot N\Bigg),$$

where:

$$C_1 = \frac{\sqrt{2\pi}}{48e^{2+1/e}} \approx 0.00489,$$
$$C_2 = \frac{1}{4e^{1+1/(2e)}} \approx 0.0765.$$

- Finally, note that by definition of $\epsilon > 0$:

$$\sum_{t=T_c+1}^{\infty} \mathbb{P}(T \geq t) \leq \sum_{t=T_c+1}^{\infty} (1-p_1)^{t-1} = \frac{1}{p_1}(1-p_1)^{T_c} \leq \frac{p_2}{p_1^2}(1-p_1)^{T_c} = \frac{1}{8}\epsilon.$$

For the multivariate version (i.e., $n > 1$), the same proof follows, albeit with the multivariate version of the Dvoretzky-Kiefer-Wolfowitz-Massart inequality. For more details, see Naaman, 2021 Naaman (2021). ■

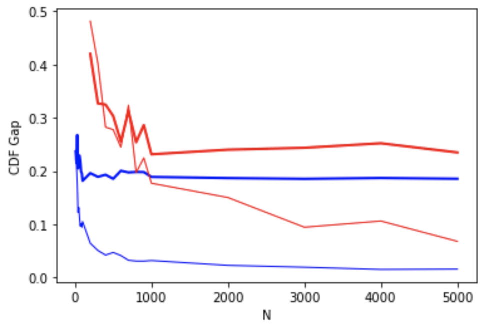


Figure 4: CDF Gap between vs. N, for the 2-link traffic network example. Here, red and blue correspond to the baseline and our method, respectively, while thick and thin lines correspond to the null and alternative hypotheses, respectively. Our approach correctly identifies the null hypothesis dataset with a relatively small number of samples, while the naive aggregation method fails to do so (thin blue curve).

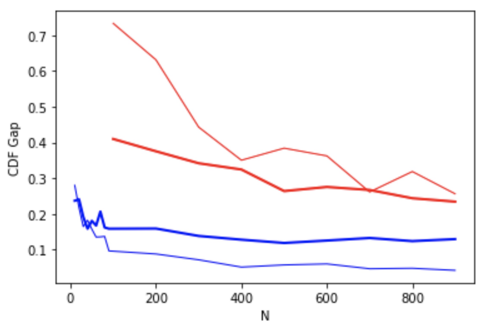


Figure 5: CDF Gap between vs. N, for the 3-link traffic network example. The color and thickness schemes are identical to those of the single-link and 2-link plots in Figures 2 and 4.

Appendix C. Experiment Results

C.1. Multi-link Traffic Networks

For the multi-link traffic network, we use the dynamics: (Maheshwari et al. (2022a))

$$x_i[t+1] = (1-\mu) \cdot x_i[t] + \mu \cdot \frac{e^{-\beta \cdot x_i[t]}}{\sum_{j=1}^{R} e^{-\beta \cdot x_j[t]}} \cdot u[t] + w[t], \qquad \forall t \in [T], i \in [R], \tag{8}$$

$$A[t] \sim \mathcal{P}(x[t]), \tag{9}$$

where $x_i[t]$ denotes the traffic flow on each link $i \in [R]$, $u[t] \in \mathbb{R}$ and $w[t] \in \mathbb{R}$, and T_h, are the input, zero-mean noise terms, and time horizon, as before. Here, we set $T = 250$, $\mu(0) = 0.3$, $\mu(1) = 0.2$, $u(t) = 100R$ for each $t \in [T]$, and we again draw $w[t]$ i.i.d. from the continuous uniform distribution on $(-10, 10)$. As with the single-link case, we created two datasets for the null and alternative hypotheses. For the null hypothesis, we fix $\mathcal{P}(x[t])$ to be Bernoulli(0.02); for the alternative hypothesis, we set $\mathcal{P}(x[t])$ to be Bernoulli(0.02) when $x[t] < 105$, and Bernoulli(0.30) when $x[t] \geq 105$. Again, this setting encodes the situation where higher traffic loads cause higher accident probabilities.

Similar to the single-link case, we compute the maximum CDF gap $\sup_{x\in\mathbb{R}} |\hat{b}_1^N(x) - \hat{b}_2^N(x)|$ as functions of N (thin lines), and the empirical CDFs of $X_{t-1}|T = t$ and X_{t-1} (thick lines) for both the null and alternative hypotheses. We again observe that our method distinguishes between the two hypotheses at a smaller sample number N compared to the baseline method.

Analogous results hold for a 3-link system with dynamics as given by (8) and are presented in Figure 5.