Journal of Information Technology Review

DLINE JOURNALS

Print ISSN: 0976-3511
Online ISSN: 0976-2922

JITR 2026: 17 (3)
https://doi.org/10.6025/jitr/2026/17/3/120-147

# Probability Concentration and Rare-Event Characterization in High-Dimensional Correlated Multivariate Bernoulli Systems

Ricardo Rodríguez Jorge
CEIT Research Center
Data Analysis and Information Management Group/ICT Division
Spain
rrodriguezj@ceit.es

## ABSTRACT

*High-dimensional correlated multivariate Bernoulli systems exhibit exponentially large state spaces, making the characterization of rare events and probability concentration a significant computational challenge. This study presents a comprehensive analytical framework to investigate sparsity, dominance, and tail behavior within such discrete binary systems. Utilizing a benchmark dataset of 20 correlated Bernoulli variables, we systematically analyze state spaces across increasing dimensionalities (4 to 20 bits) through tail probability evaluation, Herfindahl Hirschman Index (HHI) concentration metrics, activation pattern characterization, and Generalized Extreme Value (GEV) modeling.*

*Results reveal extreme sparsity, with over 99.9% of configurations at the highest dimension possessing probabilities below* $10^{-4}$ *. Despite a theoretical state space of 1,048,576 configurations, probability mass is heavily concentrated: a single dominant state captures approximately 20% of the total mass, and the top 100 states account for over 50%. Consequently, the effective number of states remains remarkably low (approximately 23.51), indicating that practical system complexity is vastly smaller than its combinatorial size. Furthermore, GEV modeling confirms that extreme configurations follow a predictable, heavy-tailed statistical regime rather than random fluctuations.*

*These findings demonstrate that high-dimensional correlated binary systems are governed by a small subset of highly activated configurations. This insight enables efficient, reduced state representations crucial for advancing reliability analysis, anomaly detection, and risk assessment in complex stochastic networks.*

**Keywords:** Correlated Multivariate Bernoulli Systems, Rare Event Characterization, Probability Concentration, High Dimensional State Space, Extreme Value Theory, Herfindahl Hirschman Index, Tail probability Analysis, Sparsity, Effective Complexity

**Received:** 5 January 2026, Revised 16 March 20026, Accepted 28 March 2026



# 1. Introduction

Extreme and rare events constitute a fundamental characteristic of many natural and engineered systems. These events correspond to large excursions that occur intermittently and often exhibit values several standard deviations above normal behavior. Such phenomena are typically associated with nonlinear interactions and energy transfers between scales or modes and have attracted considerable attention because of their profound consequences and the need for reliable prediction and statistical characterization [1].

## 2. Related Work

Rare events, characterized by extremely low occurrence probabilities but potentially catastrophic impacts, represent a major challenge in uncertainty quantification and risk analysis. Accurate estimation of these probabilities is essential in numerous safety critical applications, including infrastructure networks, nuclear power plants, aerospace systems, autonomous vehicles, and advanced manufacturing environments, where reliable operation under uncertain conditions is of paramount importance [2]. Reliability analysis in these settings primarily focuses on estimating failure probabilities, defined as the likelihood that system demands exceed available capacities. Several advanced methodologies have been developed to address this problem, including subset simulation, sequential importance sampling, and improved cross entropy approaches [3].

The significance of rare event analysis has increased considerably with the emergence of data intensive systems. In smart-city applications, particularly those involving anomaly detection and cybersecurity, rare events often contain the most valuable information embedded within massive datasets generated by the Internet of Things (IoT). Consequently, extracting and analyzing information associated with minority subsets becomes essential for effective decision-making and risk mitigation [4].

Extreme events arise across a broad spectrum of complex dynamical systems. In many chaotic systems, large excursions originate from transient instabilities that are randomly activated as the system evolves on its attractor. For instance, turbulent fluid flows exhibit persistent instabilities characterised by positive Lyapunov exponents and dissipation mechanisms, where the random activation of non normal dynamics can trigger extreme behaviour [5, 6, 7]. In nonlinear wave systems, randomness induced by dispersive phase mixing and random superposition can produce nonlinear focusing effects that lead to extreme events [8, 9].

Complex networks provide another important context in which random perturbations can induce abrupt transitions. Examples include networked populations, where stochastic effects may lead to disease extinction [10, 11], and communication systems, where perturbations can initiate the internal propagation of malware [12]. Similar mechanisms are observed in mechanical systems subjected to stochastic excitations, including parametric instabilities in ship rolling motions [13, 14] and buckling phenomena in nonlinear beams under combined axial and transverse loading conditions [15].

From a statistical perspective, considerable effort has been devoted to characterizing heavy tailed behavior and quantifying the occurrence of extreme events. Among these approaches, Extreme Value Theory (EVT) provides a rigorous framework for analyzing large deviations associated with random variables exceeding specified thresholds [16, 17]. Although conventional statistical techniques have proven successful in many applications, the increasing dimensionality and complexity of modern systems pose significant challenges for efficient probability estimation.

Recent developments have therefore focused on combining statistical reliability theory with machine learning and deep learning techniques. In particular, the reliability assessment of deep neural networks has emerged as an important research direction. The objective is to evaluate the robustness of these models under rare but critical conditions and to develop statistical methodologies capable of capturing network behavior when subjected to unusual or corrupted inputs [18].

High dimensional rare event probability estimation remains computationally challenging because the corresponding optimal importance distributions are often highly concentrated and multimodal. To address these difficulties, Tiangang Cui [19] proposed a deep importance sampling framework specifically designed for high dimensional rare event problems. The method approximates the optimal importance distribution as the pushforward of a reference distribution through a composition of order preserving transformations represented by squared tensor train decompositions [Tiangang Cui]. Building upon the deep inverse Rosenblatt transport (IRT) framework developed in [20] and [21], the approach adaptively learns complex importance densities through compositions of monotone mappings. This composite structure is particularly effective when the target distribution exhibits multiple modes and significant tail concentration. The individual transformations are constructed using functional tensor-train decomposition and cross-approximation algorithms [22, 23, 24, 25, 26].

These advances demonstrate the growing convergence between statistical reliability engineering, rare event simulation, and deep learning methodologies. [27] Nevertheless, efficient estimation of extremely small probabilities in high dimensional correlated systems remains an open challenge, motivating the development of more scalable and robust computational frameworks for rare event characterization and reliability analysis.

## 3. Research Design and Analytical Testbed Architecture

### 3.1 Overall Research Framework

The present study adopts a quantitative, exploratory research design to investigate the structure of rare events and probability concentration phenomena in high dimensional correlated multivariate Bernoulli systems. The framework combines probability theory, concentration analysis, pattern characterization, and extreme value modeling to provide a comprehensive understanding of how probability mass is distributed across a very large discrete state space.

The research process follows a sequential analytical architecture consisting of five major layers:

1. Data acquisition and preprocessing

2. State space generation and probability representation

3. Rare event and concentration analysis

4. Structural and activation pattern characterization

5. Extreme value modeling and graphical interpretation

Together, these components form an integrated computational testbed for examining sparsity, dominance, and tail behavior in correlated binary systems.

### 3.2 Analytical Testbed Architecture

The proposed testbed consists of interconnected modules designed to transform the original probability distributions into interpretable statistical representations of extreme events and concentration effects.

### Layer 1: Dataset Acquisition

The first layer receives the benchmark dataset, which contains correlated multivariate Bernoulli distributions. The dataset consists of:

- Joint probability distributions ($Q_{joint}$)
- Conditional probability distributions ($Q_{Conditional}$)
- Marginal probability distributions ($Q_{marginal}$)

Five dimensional configurations are considered:

- 4 bits
- 8 bits
- 12 bits
- 16 bits
- 20 bits

corresponding to state spaces ranging from

$$2^4=16$$

to

$$2^{20}=1,048,576$$

possible configurations.

This hierarchical structure enables investigation of how increasing dimensionality influences probability concentration and the emergence of rare states.

Layer 2: State-Space Construction and Probability Mapping
In the second layer, the joint distributions are transformed into complete binary state spaces.

Each configuration is represented as

$$x = (x_1, x_2, ....x_d), x_i \varepsilon \{0,1\}$$

where every state is associated with its corresponding occurrence probability.

This module provides the foundation for:

- Enumeration of all configurations;
- Probability ranking;
- Identification of dominant and low-probability states;
- Construction of activation-level representations.

Layer 3: Rare-Event Detection and Concentration Analysis
The third layer performs statistical characterization of rare configurations.
States satisfying

$$P(x) < \tau$$

are classified as rare events, where multiple thresholds are employed to investigate progressively extreme regions of the probability space.

This module computes:

### 3.3 Tail Probability Measures

- Percentage of states below $10^{-4}$,
- Percentage below $10^{-5}$,
- Percentage below $10^{-6}$.

### Dominant-State Analysis

Identification of:

- Maximum-probability states,
- Top-10 configurations,
- Top-100 configurations.

### Concentration Metrics

The probability concentration is quantified using the Herfindahl Hirschman Index

$$HHI = \sum_{i=1}^{N} p_i^2$$

and the effective number of states

$$N_{eff} = \frac{1}{HHI}$$

These measures provide a reduced representation of the system's practical complexity.

## Layer 4: Structural Pattern Analysis

The fourth layer investigates how probability mass is organized among binary configurations.
Several complementary analyses are performed:

### Dominant Pattern Identification

Extraction of:

- Top 20 most probable configurations;
- Bottom 20 rarest configurations.

### Marginal Probability Analysis

Examination of probability distributions across individual four bit groups.

### Activation-Level Analysis

Evaluation of the distribution of the number of active variables (number of ones).

This module reveals whether highly activated or sparsely activated patterns dominate the probability landscape.

### Layer 5: Extreme-Value Modeling

The fifth layer characterizes the statistical behavior of extreme configurations using Extreme Value Theory (EVT).

Generalized Extreme Value (GEV) analysis is employed to estimate return levels corresponding to different return periods.

The model enables:

- Quantification of rare-event magnitudes;
- Investigation of tail structure;
- Estimation of recurrence characteristics;
- Assessment of extreme-state behavior.

Return-level analysis further provides insight into the relationship between event severity and expected recurrence intervals.

### Layer 6: Visualization and Interpretation

The final layer transforms numerical outputs into graphical representations for intuitive interpretation.

Visualization components include:

- Probability distribution plots;
- Tail-distribution representations;
- Extreme-value plots;
- Return-level curves;
- Heatmaps of pattern activations.

These visual tools provide additional evidence regarding sparsity, concentration, and heavy-tailed behavior.

## 3.4 Computational Workflow

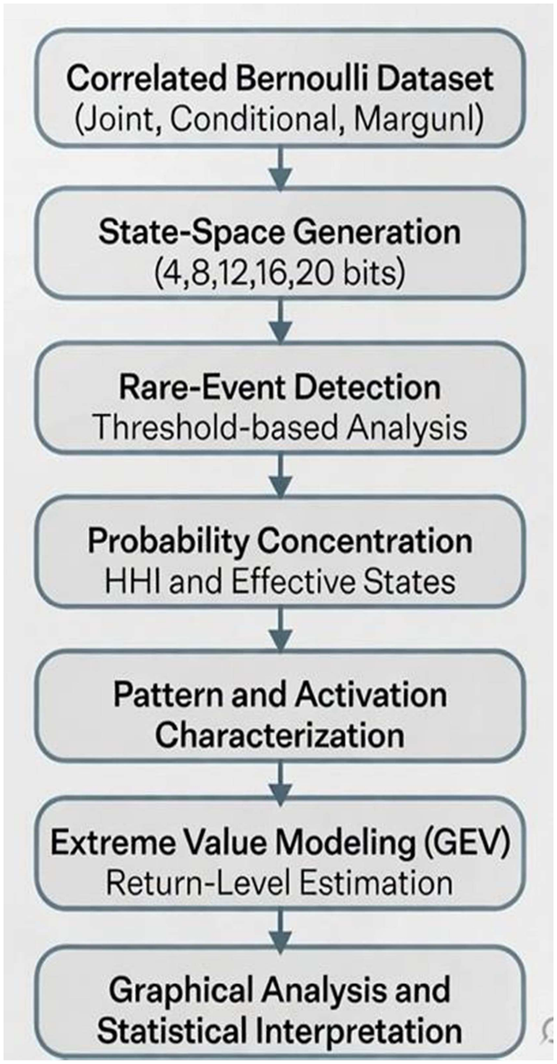
The complete analytical workflow can be summarized as

Correlated Bernoulli Dataset
↓
Probability Extraction
↓
State-Space Construction
↓
Rare-Event Identification
↓
Probability Concentration Analysis
↓
Pattern and Activation Characterization
↓

Generalized Extreme Value Modeling
↓
Graphical Analysis
↓
Interpretation of High-Dimensional Probability Structure

### 3.5 Proposed Architecture Diagram

The analytical architecture underlying the study may be represented as:




### 3.6 Research Paradigm

The proposed architecture can be viewed as a multi-layer analytical testbed integrating probability theory, concentration measures, and extreme-value analysis for characterizing rare events in high dimensional correlated multivariate Bernoulli systems. It provides a unified framework for understanding sparsity, dominant configurations, effective complexity, and tail behavior in large binary state spaces and is readily extendable to reliability engineering, machine learning, information theory, and stochastic network applications.

This section would fit naturally after Section 3.2 Methodological Framework and before the Results section, and can serve as either "Research Design," "Analytical Framework," "Testbed Architecture," or "Proposed System Architecture" in a journal paper.

# 4. Dataset and Methodology

## 4.1 Dataset Description

The study utilizes an open-access dataset comprising probability distributions of correlated multivariate Bernoulli random variables. The dataset was developed to address the limited availability of benchmark datasets representing high-dimensional correlated binary systems, which are of considerable importance in fields such as machine learning, hardware security, statistical inference, and information theory. In contrast to continuous benchmark datasets that are widely available in public repositories, resources describing correlated discrete distributions remain comparatively scarce.

Correlated binary variables arise naturally in numerous applications. In hardware security, Physically Unclonable Functions (PUFs) generate device-specific binary responses that exhibit statistical dependence. Similarly, machine learning models involving binary features require an understanding of the underlying dependency structure to accurately capture joint behavior. Correlated Bernoulli processes are also fundamental in hypothesis testing, discrete-time signal processing, source coding, and information-theoretic analyses.

To facilitate research in these areas, the dataset provides a comprehensive representation of the distributions associated with twenty correlated Bernoulli variables. The dataset is publicly available through IEEE DataPort (DOI: 10.21227/vhhe-0m96) and is distributed as a MATLAB file, Correlated_multivariate_Bernoulli.mat, with a size of approximately 3.75 MB.

### 4.1.1 Multi-Scale Representation

To support analyses across different dimensionalities, the twenty-variable system is partitioned into five progressively increasing bit lengths: 4, 8, 12, 16, and 20 bits. This hierarchical structure enables investigation of how probability distributions evolve as dimensionality increases and allows direct comparison of concentration and sparsity characteristics across different state space sizes.

The dataset is organized into three principal components containing joint, conditional, and marginal probability distributions. Together, these components provide complementary descriptions of the statistical structure of the correlated binary system.

### 4.1.2 Joint Probability Distributions

The variable (*Q_{joint}*) contains the complete joint probability distributions corresponding to the five dimensional levels. Each cell represents a specific bit length and stores the probabilities associated with all possible binary configurations.

For a system containing (d) binary variables, the number of possible states equals

[ N=2^{d}. ]

N=2^{d}”}}

Consequently, the dimensional levels considered in this study contain 16, 256, 4,096, 65,536, and 1,048,576 states, respectively.

The complete joint distributions enable direct examination of state space probabilities and provide the basis for entropy estimation, concentration analysis, and rare event characterization.

### 4.1.3 Conditional Probability Distributions

The variable (Q_{conditional}) captures dependencies among successive groups of variables. Specifically, each conditional distribution describes the probabilities of the next four bits, conditioned on previously observed sequences. For example, bits 5–8 are conditioned on bits 1–4, while bits 9–12 are conditioned on bits 1–8.

These conditional distributions provide insight into the sequential dependence structure embedded within the correlated Bernoulli process and are particularly useful for Markov modeling, sequential prediction, and machine learning applications involving autoregressive or Bayesian frameworks.

#### 4.1.4 Marginal Probability Distributions

The variable ($Q_{marginal}$) contains the marginal distributions associated with five non-overlapping four-bit blocks corresponding to bits 1–4, 5–8, 9–12, 13–16, and 17–20.

Marginal distributions allow the behavior of individual variable groups to be examined independently and facilitate entropy estimation and independence testing among different segments of the binary sequence.

#### 4.1.5 Accompanying Analytical Tool

The dataset is accompanied by a MATLAB utility, Extract_conditional_and_marginal_distributions.m, which enables systematic extraction of conditional and marginal distributions from arbitrary discrete probability models. The modular nature of this tool allows researchers to apply the same extraction framework to other correlated discrete datasets, thereby enhancing reproducibility and promoting standardized preprocessing procedures.

### 4.2 Methodological Framework

The objective of the present study is to investigate the occurrence and structural characteristics of rare joint events within a high dimensional correlated multivariate Bernoulli system. Because the variables are binary, conventional quantile based approaches commonly employed for continuous distributions are not directly applicable. Instead, rare events are defined in terms of low-probability joint configurations within the discrete state space.

#### 4.2.1 Representation of Binary Configurations

A multivariate Bernoulli system consisting of ($d$) variables generates binary vectors of the form

[$=(x_1,x_2,,x_d)$,]

where

[$x_i\{0,1\}$]

$x_i \in \{0,1\}$

The complete state space contains ($2^{d}$) possible configurations. Each configuration is associated with a probability obtained directly from the joint distribution.

Unlike continuous random variables, where extreme events are often defined via quantiles, rare events in a multivariate Bernoulli system correspond to specific binary patterns with probabilities below predefined thresholds.

#### 4.2.2 Rare Event Definition

A configuration ($x$) is classified as rare whenever its occurrence probability satisfies

[$P(x)<\tau$,

where (tau) denotes a prescribed threshold.

*P(x)<tau*

In the present analysis, several probability thresholds were considered to characterize increasingly extreme regions of the state space. In particular, thresholds corresponding to 5%, 1%, 0.5%, and 0.1% probabilities were employed to quantify the prevalence and cumulative contribution of rare configurations.

#### 4.2.3 Analytical Procedure

The analysis proceeded through a sequence of complementary stages designed to characterize both ordinary and extreme regions of the probability distribution.

First, the joint probability distributions corresponding to each dimensional level were extracted from the dataset and converted into a state space representation. This allowed probabilities associated with every binary configuration to be enumerated explicitly.

Second, probability distributions were examined to identify dominant configurations and low-probability states. Configurations whose probabilities fell below selected thresholds were classified as rare events, and their collective contribution to the total probability mass was evaluated.

Third, concentration properties were investigated through measures such as the Herfindahl Hirschman Index and the effective number of states. These metrics provide information regarding the extent to which probability mass is concentrated among a limited number of configurations.

Fourth, the structure of dominant and rare configurations was examined through pattern analysis, marginal probability evaluation, and activation level distributions. These analyses provide insight into the mechanisms responsible for the emergence of highly probable states.

Finally, extreme-value theory was employed to investigate the tail behavior of the distribution. Generalized Extreme Value modeling and return level analysis were used to quantify the magnitude and recurrence characteristics of rare events and to determine whether extreme configurations follow a structured statistical regime.

### 4.3 Visualization and Statistical Characterization

To facilitate interpretation of the probability landscape, several graphical representations were employed. Probability distribution plots were used to visualize sparsity and concentration effects, whereas heatmaps and activation distributions provided insight into the structural organization of binary configurations. Extreme-value plots and return level curves were further utilized to characterize the upper tail behavior of the distribution.

In addition to graphical analysis, statistical characterization was performed using entropy measures, concentration indices, cumulative probability mass calculations, and effective state measures. These complementary approaches provide a comprehensive description of the complexity, sparsity, and concentration properties of the correlated multivariate Bernoulli system.

### 4.4 Potential Application Domains

The analytical framework developed in this study is applicable to a broad range of research areas involving correlated binary processes. In hardware security, the methodology can be employed to investigate Physically Unclonable Functions and evaluate the reliability of authentication mechanisms. In machine learning, the framework provides benchmark data for assessing generative models and dependency learning algorithms. Applications also extend to hypothesis testing, discrete time signal processing, information theory, and stochastic modeling, where understanding the structure of high dimensional binary distributions is of fundamental importance.

Overall, the dataset and methodology provide a unified framework for investigating sparsity, concentration phenomena, and extreme events in correlated multivariate Bernoulli systems. By combining complete joint distributions with rare-event analysis and extreme value modeling, the proposed framework enables systematic exploration of the effective complexity and probabilistic.

## 5. Rare-Event Characterization and Probability Concentration Analysis

The correlated multivariate Bernoulli system was examined to investigate how probability distributions evolve as dimensionality increases. Because the number of possible configurations grows exponentially with the number of binary variables, high dimensional systems are expected to exhibit substantial sparsity and strong concentration effects. To characterize these properties, four complementary analyses were performed: tail probability analysis, identification of dominant configurations, risk concentration measurement using the Herfindahl Hirschman Index (HHI), and evaluation of cumulative probability mass among the most probable states. Together, these analyses provide a comprehensive description of sparsity, concentration, and rare-event behavior within the multivariate state space.

### 5.1 Tail Probability Analysis

Tail probability analysis was conducted to determine the proportion of configurations whose probabilities fall below selected thresholds. As dimensionality increases, the number of possible states expands exponentially, causing an increasing fraction of the state space to consist of extremely low probability events.

| Distribution Level | Number of States | $P < 10^{-6}$ | $P < 10^{-5}$ | $P < 10^{-4}$ |
|---|---|---|---|---|
| Level 1 | 16 | 0.0000 | 0.0000 | 0.0000 |
| Level 2 | 256 | 0.0000 | 0.0000 | 0.0000 |
| Level 3 | 4,096 | 0.0000 | 0.0156 | 0.7166 |
| Level 4 | 65,536 | 0.7525 | 0.9362 | 0.9892 |
| Level 5 | 1,048,576 | 0.9929 | 0.9986 | 0.9996 |

Table 1. Tail Probability Analysis

The results reveal a pronounced increase in the proportion of extremely small probabilities as the system's dimensionality increases. At Levels 1 and 2, virtually no states fall below the selected thresholds. Beginning at Level 3, however, a substantial fraction of configurations enter the tail region, and this behavior becomes increasingly dominant at higher dimensions. At Level 5, approximately 99.96% of all states possess probabilities below $10^{-4}$, indicating that only a very small subset of configurations contributes appreciably to the total probability mass.

These findings demonstrate that the probability landscape becomes highly sparse as dimensionality increases. Although the number of theoretically possible states grows rapidly, most configurations become practically negligible, while a relatively small subset governs the overall behavior of the system.

### 5.2 Extreme-State Identification

To identify the configurations that dominate the distribution, the state with the largest probability was determined at each dimensional level.

| Level | Most Likely State Index | Maximum Probability |
|---|---|---|
| 1 | 15 | 0.5014 |
| 2 | 255 | 0.3711 |
| 3 | 4095 | 0.2930 |
| 4 | 65535 | 0.2364 |
| 5 | 1048575 | 0.2031 |

Table 2. Extreme-State Identification

The maximum probabilities corresponding to these dominant states are summarized below.

| Level | Maximum Probability |
|---|---|
| 1 | 50.14% |
| 2 | 37.11% |
| 3 | 29.30% |
| 4 | 23.64% |
| 5 | 20.31% |

Across all dimensional levels, the most probable configuration corresponds to the final indexed state of the joint distribution. Although the probability associated with this state decreases gradually as the number of configurations increases, the dominant state retains a remarkably large share of the overall probability mass.

Even at Level 5, where the system contains 1,048,576 possible configurations, a single state still accounts for approximately one fifth of the total probability mass. This persistence of dominant states demonstrates that increasing dimensionality does not eliminate concentration effects. Instead, system behavior remains strongly influenced by a small number of highly probable configurations.

### 5.3 Risk Concentration Analysis

The degree of probability concentration was quantified using the Herfindahl Hirschman Index (HHI):

[ HHI=/sum_{i=1}{N}p_i^2 ]

where (*p_i*) denotes the probability associated with state (i).

The reciprocal of the HHI defines the effective number of states:

N_{\mathrm{eff}}=\frac{1}{HHI}

N_{\mathrm{eff}}=\frac{1}{HHI}

This measure represents the number of equally probable states required to generate the observed level of concentration.

| Level | HHI | Effective Number of States (1/HHI) |
|---|---|---|
| 1 | 0.2727 | 3.67 |
| 2 | 0.1443 | 6.93 |
| 3 | 0.0891 | 11.23 |
| 4 | 0.0579 | 17.27 |
| 5 | 0.0425 | 23.51 |

Table 3. Risk Concentration Measures

The HHI decreases progressively as dimensionality increases, indicating that probability mass becomes distributed across a larger number of configurations. Nevertheless, the effective number of states remains extremely small relative to the total size of the state space.

For the Level 5 system, despite the existence of 1,048,576 possible configurations, the effective number of states equals only 23.51. Consequently, the system behaves as though the entire probability distribution were concentrated among only a few dozen representative states. This result indicates that the system's practical complexity is substantially smaller than its theoretical combinatorial complexity.

### 5.4 Concentration of Probability Mass

To further quantify concentration effects, the cumulative probability mass contained within the most probable configurations was evaluated.

| Level | Top-1 State | Top-10 States | Top-100 States |
|---|---|---|---|
| 1 | 0.5014 | 0.8865 | 1.0000 |
| 2 | 0.3711 | 0.5846 | 0.9417 |
| 3 | 0.2930 | 0.4381 | 0.7608 |
| 4 | 0.2364 | 0.3497 | 0.6112 |
| 5 | 0.2031 | 0.2912 | 0.5076 |

Table 4. Concentration of Probability Mass

The concentration analysis reveals a highly uneven probability structure. At Level 5, the single most probable state captures 20.31% of the total probability mass, whereas the ten most probable states account for 29.12%. More strikingly, the 100 most probable configurations account for 50.76% of the entire distribution.

Considering that the system contains 1,048,576 possible states, only one hundred configurations approximately 0.0095% of all states account for more than half of the probability mass. This observation highights an extreme concentration phenomenon and confirms that system behavior is governed primarily by a very limited subset of configurations.

### 5.5 Summary of Rare Event Characteristics

The rare event analysis demonstrates that increasing dimensionality yields a highly sparse, strongly concentrated probability landscape. Although the number of possible configurations grows exponentially, the associated probability mass becomes increasingly localized among a small set of dominant states.

Three major characteristics emerge from the analysis. First, the state space exhibits extreme sparsity, with more than 99.9% of configurations at the highest dimensional level having probabilities below $10^{-4}$. Second, dominant states persist even in very large systems, with the most probable configuration retaining approximately 20% of the total probability mass. Third, strong concentration effects are evident:

the effective number of states remains only 23.51 despite more than one million possible configurations, and the top 100 states contain over half of the total probability mass.

Collectively, these findings indicate that high dimensional correlated Bernoulli systems possess heavy tailed occupancy structures, pronounced risk concentration, and substantial sparsity effects, making their behavior fundamentally different from that of uniform or weakly concentrated probability distributions.

## 6. Descriptive Characteristics of the High-Dimensional State Space

Following the rare-event characterization, additional analyses were performed to examine the internal structure of the 20-dimensional correlated multivariate Bernoulli system. These analyses provide insight into the overall statistical properties of the state space, the prevalence of rare configurations, the dominance of common patterns, marginal behavior of individual states, and the distribution of activation levels across binary configurations. Collectively, these results reveal how probability mass is organized within the high dimensional system and provide further evidence of strong concentration and sparsity effects.

### 6.1 Overall Statistical Characteristics

The fundamental properties of the generated state space are summarized in Table 5.

| Observations (States) | Variables | Entropy (bits) | Minimum Probability | Maximum Probability |
|---|---|---|---|---|
| 1,048,576 | 20 | 12.70768 | 0 | 0.203119 |

Table 5. Basic Statistics

The state space consists of 1,048,576 unique configurations generated by 20 binary variables, resulting in an extremely large combinatorial space. The estimated entropy of 12.71 bits indicates substantial uncertainty and diversity within the system. However, this value remains considerably below the theoretical maximum entropy of 20 bits expected under a uniform distribution.

The discrepancy between the observed and theoretical entropy values indicates that the probability mass is concentrated in a relatively small subset of configurations. Furthermore, while the maximum state probability is 0.2031, many states have probabilities effectively zero, confirming a highly sparse probability landscape. These observations indicate that the distribution is far from uniform and is instead characterized by strong concentration effects.

### 6.2 Rare Event Summary

To quantify the prevalence of low-probability configurations, states were classified using multiple probability thresholds.

| Threshold | Count | Percent States | Probability Mass |
|---|---|---|---|
| 0.05 | 1,048,575 | 99.9999 | 0.796881 |
| 0.01 | 1,048,572 | 99.99962 | 0.758260 |
| 0.005 | 1,048,562 | 99.99866 | 0.685055 |
| 0.001 | 1,048,541 | 99.99666 | 0.637123 |

Table 7. Rare Events Summary

The results demonstrate that nearly all configurations belong to the rare-event category, even when relatively high probability thresholds are considered. At the 5% threshold, approximately 99.9999% of states are classified as rare and collectively account for 79.69% of the total probability mass. Even when the threshold is reduced to 0.1%, more than 99.996% of all configurations remain classified as rare while still containing over 63% of the overall probability.

These findings reveal an important characteristic of the distribution. Although individual rare states contribute negligibly on their own, their collective contribution remains substantial. Consequently, the state space is numerically dominated by rare configurations, a property commonly associated with heavy-tailed and high-dimensional stochastic systems.

### 6.3 Dominant Binary Configurations

To identify the configurations that contribute most strongly to the overall distribution, the twenty most probable patterns were extracted.

| Rank | Pattern | Probability | Sum Ones |
|---|---|---|---|
| 1 | 1.11E+19 | 0.203119 | 20 |
| 2 | 1.11E+19 | 0.015012 | 19 |
| 3 | 1.11E+19 | 0.012681 | 19 |
| 4 | 1.11E+19 | 0.010929 | 19 |
| 5 | 1.11E+19 | 0.009863 | 19 |
| 6 | 1.11E+19 | 0.008618 | 19 |
| 7 | 1.01E+19 | 0.008495 | 19 |

| | | | |
|---|---|---|---|
| 8 | 1.11E+19 | 0.008009 | 19 |
| 9 | 1.11E+19 | 0.007765 | 19 |
| 10 | 1.1E+19 | 0.006748 | 19 |
| 11 | 1.11E+19 | 0.006522 | 19 |
| 12 | 1.11E+19 | 0.006510 | 19 |
| 13 | 1.11E+19 | 0.005427 | 19 |
| 14 | 1.11E+18 | 0.005248 | 19 |
| 15 | 1.11E+19 | 0.004994 | 19 |
| 16 | 1.11E+19 | 0.004611 | 19 |
| 17 | 1.11E+19 | 0.004483 | 19 |
| 18 | 1.11E+19 | 0.004113 | 19 |
| 19 | 1.11E+19 | 0.004111 | 19 |
| 20 | 1.11E+19 | 0.003830 | 19 |

Table 8. Top 20 Most Common Patterns

The most probable configuration possesses a probability of 0.2031 and corresponds to the state in which all twenty variables are active. Most of the remaining dominant patterns contain nineteen active variables, indicating that highly activated configurations dominate the probability structure.

The sharp decline in probability between the first-ranked configuration and subsequent states further demonstrates the existence of a highly concentrated distribution. These results suggest that the system strongly favors densely populated states rather than balanced or sparsely activated configurations.

### 6.4 Rarest Configurations

The least probable states were also examined to characterize the lower tail of the distribution.

| **Rank** | **Pattern** | **Probability** | **Sum Ones** |
|---|---|---|---|
| 1 | 1E+18 | 9.54E-08 | 10 |
| 2 | 1E+18 | 9.54E-08 | 10 |
| 3 | 1E+18 | 9.54E-08 | 11 |
| 4 | 1E+18 | 9.54E-08 | 9 |

| | | | |
|---|---|---|---|
| 5 | 1E+18 | 9.54E-08 | 10 |
| 6 | 1E+18 | 9.54E-08 | 10 |
| 7 | 1E+18 | 9.54E-08 | 12 |
| 8 | 1E+18 | 9.54E-08 | 9 |
| 9 | 1E+18 | 9.54E-08 | 10 |
| 10 | 1E+18 | 9.54E-08 | 11 |
| 11 | 1E+18 | 9.54E-08 | 11 |
| 12 | 1E+18 | 9.54E-08 | 10 |
| 13 | 1E+18 | 9.54E-08 | 12 |
| 14 | 1E+18 | 9.54E-08 | 9 |
| 15 | 1E+18 | 9.54E-08 | 10 |
| 16 | 1E+18 | 9.54E-08 | 9 |
| 17 | 1E+18 | 9.54E-08 | 9 |
| 18 | 1E+18 | 9.54E-08 | 8 |
| 19 | 1E+18 | 9.54E-08 | 9 |
| 20 | 1E+18 | 9.54E-08 | 10 |

Table 9. Bottom 20 Rarest Patterns

The rarest configurations exhibit probabilities on the order of 10{ w , which are several orders of magnitude smaller than those associated with the dominant states. Unlike the most common patterns, these configurations generally contain between eight and twelve active variables.

The contrast between Tables 8 and 9 highlights the pronounced asymmetry of the probability distribution. Whereas highly activated states dominate the upper tail, moderately activated states occupy the extreme lower tail and contribute negligibly to the overall probability mass.

### 6.5 Rare Patterns Below the One Percent Threshold

Table 10. Rare Patterns Under 1% Probability

*(Available as Appendix files.)*

Table 10 provides detailed information concerning configurations whose occurrence probabilities fall below the 1% threshold. Examination of these states offers additional insight into the structural characteristics of rare events and provides further evidence supporting the heavy tailed nature of the distribution. Since these patterns constitute the majority of the state space, they reinforce the conclusion that only a very small fraction of configurations contribute substantially to the overall probability mass.

### 6.6 Marginal Probability Structure

The marginal probabilities associated with individual states across the five distribution levels are presented in Table 11.

| State | Level1 | Level2 | Level3 | Level4 | Level5 |
|---|---|---|---|---|---|
| 0 | 0.039249 | 0.042268 | 0.043354 | 0.029192 | 0.036287 |
| 1 | 0.017627 | 0.020715 | 0.02015 | 0.024609 | 0.021191 |
| 2 | 0.031367 | 0.022786 | 0.025048 | 0.020896 | 0.018082 |
| 3 | 0.027552 | 0.022224 | 0.022608 | 0.02814 | 0.025064 |
| 4 | 0.016371 | 0.015859 | 0.015754 | 0.018092 | 0.022759 |
| 5 | 0.020048 | 0.020015 | 0.022265 | 0.025158 | 0.027411 |
| 6 | 0.02316 | 0.024466 | 0.02169 | 0.026191 | 0.020039 |
| 7 | 0.054295 | 0.061396 | 0.058925 | 0.057413 | 0.057098 |
| 8 | 0.017071 | 0.022236 | 0.021148 | 0.014043 | 0.017538 |
| 9 | 0.019641 | 0.026062 | 0.022072 | 0.025676 | 0.019643 |
| 10 | 0.024169 | 0.02273 | 0.025584 | 0.022057 | 0.024403 |
| 11 | 0.052966 | 0.050705 | 0.049772 | 0.064774 | 0.066831 |
| 12 | 0.022756 | 0.020508 | 0.020194 | 0.022772 | 0.025552 |
| 13 | 0.07683 | 0.061986 | 0.064785 | 0.070087 | 0.059229 |
| 14 | 0.055544 | 0.058869 | 0.057018 | 0.076346 | 0.064204 |
| 15 | 0.501353 | 0.507175 | 0.509633 | 0.474554 | 0.494669 |

Table 11. Marginal Probabilities

Inspection of the marginal distributions reveals that most variables exhibit relatively small probabilities, generally ranging between approximately 0.015 and 0.076. In contrast, State 15 consistently exhibits probabilities approaching 0.50 across all dimensional levels.

The persistence of this dominant marginal component suggests that State 15 plays a central role in shaping the joint probability structure and contributes substantially to the emergence of highly probable configurations. Furthermore, the relative stability of the marginal probabilities across levels indicates that the overall architecture of the probability distribution remains structurally consistent despite increases in dimensionality.

## 6.7 Distribution of the Number of Active Variables

To investigate activation characteristics within the state space, the probability distribution of the total number of active variables was examined.

| Num_ Ones | Probability |
|---|---|
| 0 | 0.002351 |
| 1 | 0.004758 |
| 2 | 0.007031 |
| 3 | 0.009317 |
| 4 | 0.011676 |
| 5 | 0.014105 |
| 6 | 0.016697 |
| 7 | 0.01954 |
| 8 | 0.022527 |
| 9 | 0.025844 |
| 10 | 0.029666 |
| 11 | 0.033843 |
| 12 | 0.038827 |
| 13 | 0.044455 |
| 14 | 0.05142 |
| 15 | 0.060029 |
| 16 | 0.070974 |
| 17 | 0.085387 |
| 18 | 0.106827 |
| 19 | 0.141607 |
| 20 | 0.203119 |

Table 12. Distribution of the Number of Active States

The probability associated with a configuration increases monotonically with the number of active variables. The highest probability occurs when all twenty variables are active, accounting for 20.31% of the total probability mass. Configurations containing nineteen active variables represent the second most probable class, contributing 14.16%.

Conversely, states containing only a small number of active variables occur with substantially lower probabilities. This monotonic behavior indicates a pronounced positive skew toward highly activated configurations and explains the concentration of probability mass observed among the dominant states identified earlier.

### 6.8 Summary of State-Space Characteristics

The descriptive analyses consistently demonstrate that the 20-dimensional correlated Bernoulli system exhibits a highly structured and strongly concentrated probability landscape. Although the system contains more than one million possible configurations, probability mass is distributed very unevenly. Highly activated states dominate the upper tail of the distribution, while the overwhelming majority of configurations occur with extremely small probabilities.

The coexistence of dominant patterns, extensive rare-event populations, and stable marginal structures suggests that the system possesses strong dependence characteristics and substantial asymmetry. These properties collectively reinforce the conclusions obtained from the rare-event analyses and indicate that the effective behavior of the system is governed by a comparatively small number of highly probable configurations despite its enormous combinatorial complexity.

## 7. Extreme Value Modeling and Graphical Analysis

The preceding analyses demonstrated that the correlated multivariate Bernoulli system possesses a highly concentrated probability structure characterized by extensive sparsity and strong dominance effects. To further investigate the behavior of extreme configurations, extreme-value theory was employed to quantify the magnitude and recurrence characteristics of rare events. In addition, graphical analyses were used to visualize the distributional properties and concentration patterns identified in the numerical results.

### 7.1 Generalized Extreme Value Analysis

Extreme value analysis provides a framework for describing the behavior of the tails of probability distributions and estimating the recurrence characteristics of rare events. In the present study, the Generalized Extreme Value (GEV) framework was used to estimate return levels corresponding to different return periods.

| **Return Period** | **Return Level** |
|---|---|
| 2 | 1.67-07 |
| 5 | 6.01-07 |
| 10 | 1.29-06 |
| 20 | 2.62-06 |
| 50 | 6.48-06 |
| 100 | 1.27-05 |

Table 13. GEV Return Levels

The estimated return levels increase systematically with increasing return periods, indicating that increasingly extreme events correspond to progressively larger thresholds. For example, the return level rises from approximately (1.67\times10^{-7}) for a two-period interval to (1.27\times10^{-5}) for a one hundred period interval. This monotonic behavior is consistent with theoretical expectations from extreme value theory and demonstrates the existence of a well-defined tail structure within the distribution.

The results further suggest that extreme configurations follow a predictable pattern rather than arising purely from random fluctuations. Consequently, the GEV framework provides a useful mechanism for quantifying the severity and recurrence characteristics of rare states within high dimensional correlated Bernoulli systems.

### 7.2 Probability Distribution Characteristics

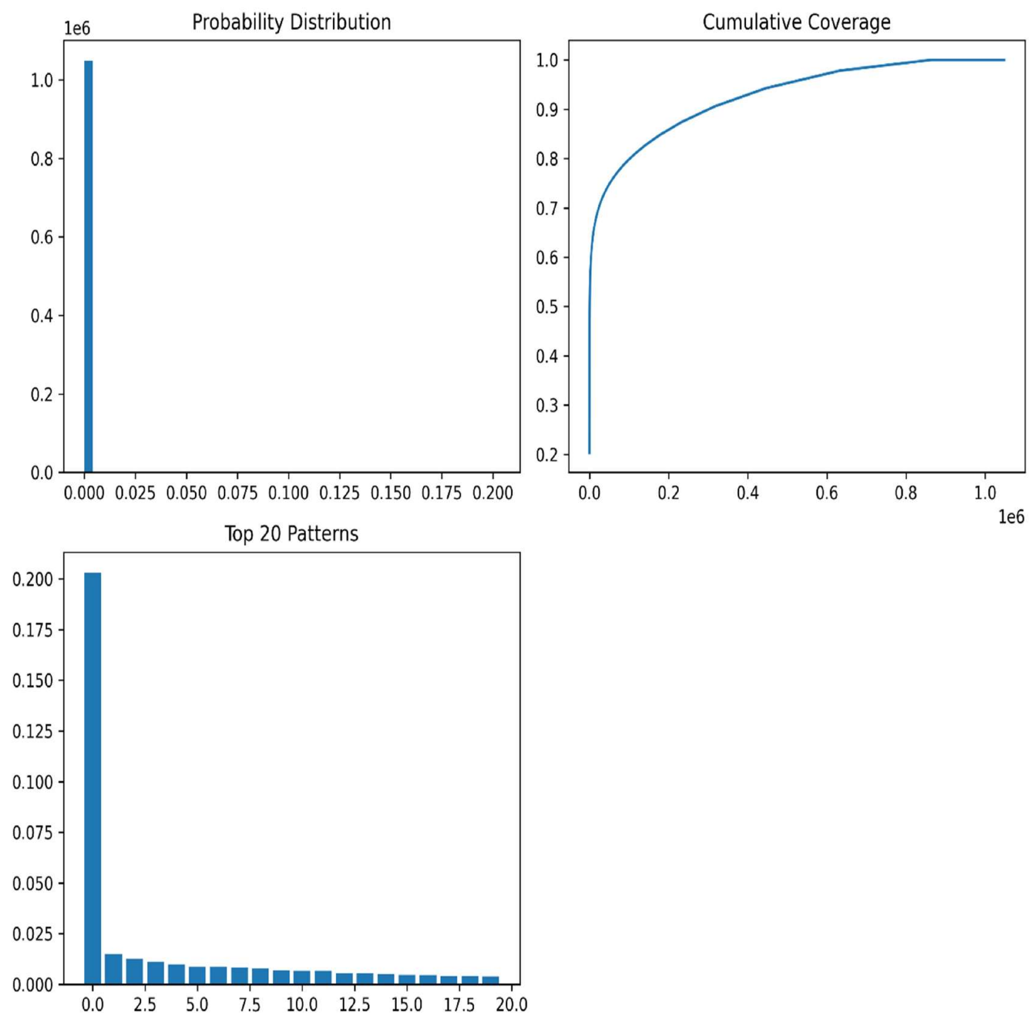



Figure 1. Probability Distribution

Figure 1 illustrates the overall distribution of state probabilities. The distribution exhibits pronounced right-skewness, with a very small number of configurations having relatively high probabilities, while the overwhelming majority of states have probabilities approaching zero. This asymmetric structure confirms strong sparsity and shows that probability mass is concentrated in only a limited subset of the state space.

The graphical evidence is consistent with the numerical results obtained from the tail probability and concentration analyses, further supporting the conclusion that the effective behavior of the system is governed by a comparatively small number of dominant configurations.

## 7.3 Extreme-Value Characteristics

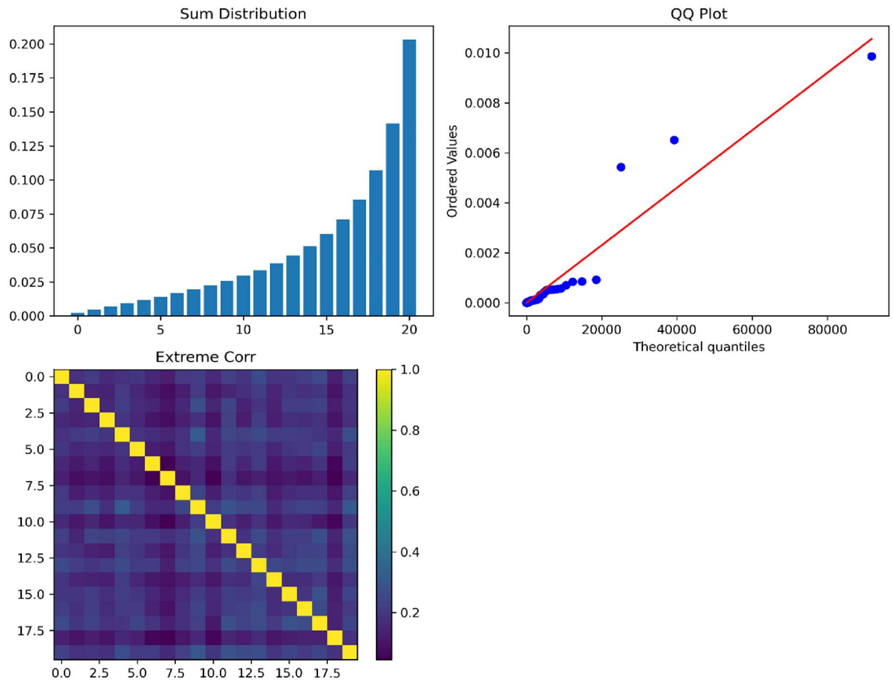



Figure 2. Extreme Value Analysis

Figure 2 illustrates the upper-tail behavior of the state probability distribution and highlights the separation between ordinary and extreme configurations. The fitted extreme value model captures the tail structure effectively, indicating that the distribution exhibits a stable, well defined extreme event regime.

The existence of such a regime suggests that rare events are not randomly scattered throughout the state space but instead follow identifiable statistical patterns. Consequently, extreme value theory provides an appropriate framework for characterizing the occurrence and magnitude of low probability configurations within the correlated Bernoulli system.

## 7.4 Return-Level Behavior

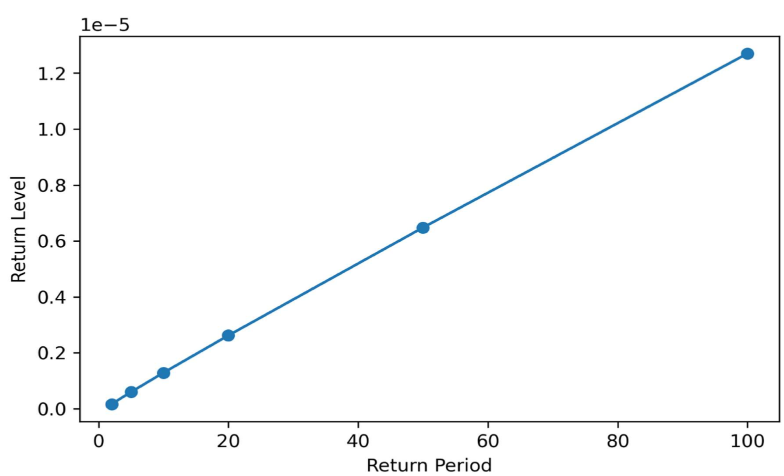



Figure 3. Return Level Plot

presents the relationship between return levels and return periods. The upward trajectory of the curve demonstrates that increasingly extreme events are associated with progressively longer recurrence intervals. The approximately monotonic pattern observed in the figure is consistent with the theoretical properties of the Generalized Extreme Value model.

The return-level plot complements the numerical estimates reported in Table 13 by providing a visual representation of the relationship between event rarity and expected magnitude. Together, these results confirm the existence of a structured and predictable tail behavior within the distribution.

### 7.5 Pattern Activation Structure

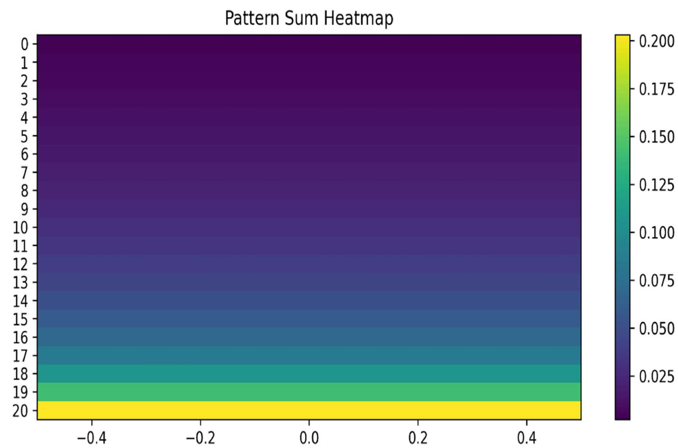



Figure 4. Pattern Sum Heatmap

Figure 4 provides a heatmap representation of the relationship between binary patterns and the number of active variables. Dense regions of the heatmap are concentrated near the upper end of the activation spectrum, indicating that configurations with many active variables account for most of the probability mass.

Conversely, configurations with relatively few active variables appear in sparsely populated regions and contribute substantially less to the overall distribution. The heatmap therefore provides intuitive confirmation of the dominance of highly activated states identified previously through the analysis of common patterns and activation distributions.

## 8. Integrated Discussion

The analyses presented throughout this study collectively reveal that the high dimensional correlated multivariate Bernoulli system exhibits a probability structure that differs fundamentally from that expected under a uniform or weakly dependent model. Although the number of possible configurations increases exponentially with dimensionality, the corresponding probability mass becomes progressively concentrated among a remarkably small subset of states.

Tail probability analysis demonstrated that increasing dimensionality yields an extremely sparse state space, in which the overwhelming majority of configurations have negligible probability. At the highest dimensional level, more than 99.9% of states fall below the (10^{-4}) probability threshold, indicating that only a minute fraction of the available configurations contribute meaningfully to the overall system behavior.

The analysis of dominant states further showed that concentration effects persist even in systems containing more than one million possible configurations. Despite the enormous combinatorial complexity, a single state continues to capture approximately twenty percent of the total probability mass, while the hundred most probable states collectively account for more than half of the entire distribution. These findings indicate that the practical complexity of the system is dramatically smaller than its theoretical state space size.

Risk concentration measures provide additional support for this conclusion. Although the state space contains 1,048,576 possible configurations, the effective number of states derived from the Herfindahl Hirschman Index equals only 23.51. Consequently, the system behaves as though probability mass were distributed among only a few dozen representative states. Such concentration effects are characteristic of strongly dependent stochastic systems and imply that a relatively small number of configurations govern the macroscopic behavior of the entire system.

Examination of common and rare patterns revealed a pronounced asymmetry between highly probable and extremely improbable states. Dominant configurations are characterized by large numbers of active variables, whereas many low-probability states contain moderate activation levels and contribute negligibly to the overall probability distribution. The monotonic increase in probability with increasing numbers of active variables further emphasizes the tendency of the system to favor densely activated configurations.

Extreme-value modeling demonstrated that the upper tail of the probability distribution possesses a well-defined structure that can be effectively characterized using the Generalized Extreme Value framework. The systematic increase in return levels with increasing return periods confirms that the occurrence of rare events follows predictable statistical behavior. The corresponding graphical analyses provide visual evidence supporting the numerical findings and illustrate the strong concentration and heavy-tailed characteristics of the distribution.

Taken together, the results indicate that the correlated multivariate Bernoulli system exhibits three defining characteristics: extreme sparsity, persistent dominance of a small number of configurations, and strong probability concentration. These properties imply that the effective dynamics of very high-dimensional binary systems are governed not by the full combinatorial state space but rather by a relatively small subset of representative configurations. Consequently, despite their enormous theoretical complexity, such systems possess an underlying structure that can be effectively described using concentration measures and extreme-value theory.

The findings provide important insights into the behavior of correlated binary systems and demonstrate that high-dimensional dependence can generate highly uneven occupancy structures characterized by heavy tails, substantial risk concentration, and remarkably low effective complexity relative to the size of the theoretical state space.

## 9. Conclusion and Future Research Directions

### 9.1 Conclusion

This study investigated the probabilistic behavior of a high dimensional correlated multivariate Bernoulli system through a comprehensive analysis of rare events, probability concentration, dominant configurations, and extreme-value characteristics. By combining tail probability analysis, concentration measures, activation distributions, pattern analysis, and Generalized Extreme Value modeling, the study provides a detailed characterization of the structural properties of large binary state spaces.

The results demonstrate that increasing dimensionality yields a highly sparse, strongly concentrated probability landscape. Although the number of possible configurations grows exponentially, the system's effective behavior is governed by a comparatively small subset of states. More than 99.9% of configurations at the highest dimensional level possess extremely small probabilities, while a single dominant state accounts for approximately 20% of the total probability mass. Furthermore, the effective number of states remains remarkably low compared with the theoretical state space size, indicating that practical complexity is substantially smaller than the combinatorial complexity.

Analysis of activation patterns revealed a strong preference for highly activated configurations, with probability mass concentrated on states with many active variables. The examination of common and rare patterns further demonstrated pronounced asymmetry between dominant and low probability configurations. In addition, extreme value modeling confirmed the presence of a structured upper tail regime and showed that rare events exhibit predictable recurrence behavior that can be effectively described using the Generalized Extreme Value framework.

Collectively, these findings establish that correlated multivariate Bernoulli systems exhibit three fundamental characteristics: extreme sparsity, persistent dominance by a limited number of configurations, and strong concentration of probability. These properties distinguish such systems from uniform or weakly dependent distributions and indicate that large scale binary systems possess an underlying structure that is considerably simpler than suggested by their theoretical dimensionality.

### 9.2 Main Contributions

The study makes several contributions to the understanding of high-dimensional correlated binary systems.

1. It provides a unified framework for analyzing sparsity and concentration phenomena in multivariate Bernoulli distributions through the integration of tail probability analysis, effective state measures, and probability mass concentration metrics.

2. It demonstrates that enormous combinatorial state spaces can exhibit remarkably low effective complexity, thereby revealing the existence of dominant representative configurations.

3. It shows that highly activated states play a central role in shaping the probability structure of correlated binary systems.

4. It establishes the applicability of extreme value theory for modeling the upper-tail behavior and recurrence characteristics of rare configurations.

5. It provides both numerical and graphical evidence that high dimensional dependence generates heavy-tailed occupancy structures and substantial risk concentration.

These contributions extend the understanding of probability concentration in correlated discrete systems and provide analytical tools applicable to a broad range of high dimensional stochastic models.

### 9.3 Practical Implications

The findings have important implications for the analysis and modeling of complex binary systems encountered in diverse application domains. Because only a relatively small number of configurations govern system behavior, efficient approximation methods and reduced state representations may provide accurate descriptions without requiring exhaustive enumeration of the entire state space.

The observed concentration phenomena are particularly relevant to applications involving reliability analysis, risk assessment, anomaly detection, network systems, biological processes, information theory, and machine learning. In such settings, understanding the structure of dominant states and extreme events can facilitate improved prediction, simulation, and decision making.

Furthermore, the existence of well defined extreme value characteristics suggests that rare event probabilities and recurrence behavior can be estimated systematically, providing valuable information for systems in which extreme configurations are associated with critical outcomes.

### 9.4 Limitations

Despite the insights obtained, several limitations should be acknowledged.

First, the analysis was conducted using a 20-dimensional correlated multivariate Bernoulli system. Although the results reveal clear concentration patterns, systems with substantially larger dimensions may exhibit additional structural properties that warrant further investigation.

Second, the study primarily focused on static probability distributions and did not consider temporal evolution or dynamic dependencies among states. Consequently, transition behavior and time varying concentration effects remain outside the scope of the present work.

Third, the analysis employed aggregate concentration measures and extreme value models without explicitly investigating the influence of different dependence structures or correlation strengths. Variations in these parameters may produce alternative probability landscapes and deserve further examination.

Finally, the conclusions are based on simulated probability structures and should be complemented by empirical studies involving real-world binary systems to assess the generalizability of the observed phenomena.

### 9.5 Future Research Directions

Several avenues for future research emerge from the present study.

Future investigations may extend the analysis to higher dimensional systems with hundreds or thousands of variables to examine the scalability of concentration phenomena and effective state behavior. Comparative studies involving different dependence structures and correlation models would provide additional insight into the mechanisms responsible for the emergence of dominant configurations.

Dynamic extensions of the framework could incorporate temporal processes through Markov chains, hidden-state models, or stochastic network representations to explore the evolution of probability concentration over time. Such approaches would enable the study of persistence, transitions, and stability within high-dimensional binary systems.

Another promising direction involves applying network science and graph theoretic methods to characterize relationships among configurations and to identify communities or hierarchical structures within the state space. Similarly, information theoretic measures and entropy based approaches may provide a deeper understanding of complexity and dependence mechanisms.

The integration of machine learning techniques with extreme value theory also represents an important avenue for future work. Combining probabilistic modeling with data driven methods may improve the identification, prediction, and characterization of rare configurations in large scale systems.

Finally, applying the proposed analytical framework to empirical datasets from fields such as communication networks, biological systems, reliability engineering, financial risk analysis, and artificial intelligence would enable validation of the theoretical findings and facilitate the development of practical methodologies for understanding high dimensional correlated binary phenomena.

In summary, the present study demonstrates that high-dimensional correlated multivariate Bernoulli systems possess an unexpectedly low effective complexity despite their enormous combinatorial size. The coexistence of extreme sparsity, strong concentration, and heavy-tailed behavior suggests that such systems can be understood in terms of a relatively small set of dominant configurations, providing a foundation for future theoretical developments and practical applications in the analysis of complex stochastic systems.

## References

[1] Sapsis, T. P. (2018). New perspectives for the prediction and statistical quantification of extreme events in high-dimensional dynamical systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2127), 20170133. https://doi.org/10.1098/rsta.2017.0133.

[2] Eshra, E. (2025). Theoretical and computational advancements for importance sampling based rare event uncertainty quantification for high dimensional safety critical systems [Doctoral dissertation, The Pennsylvania State University].

[3] Kinnear, H. J. (2025). *Niching strategies for multimodal rare-event simulation* [Doctoral thesis, University College London].

[4] She, R., Liu, S., Wan, S., Xiong, K., Fan, P. (2019). Importance of small probability events in big data: Information measures, applications, and challenges. *IEEE Access, 7*, 100363–100382.

[5] Platt, N., Sirovich, L., Fitzmaurice, N. (1991). An investigation of chaotic Kolmogorov flows. *Physics of Fluids A, 3*, 681–696. https://doi.org/10.1063/1.858074.

[6] Chandler, J., Kerswell, R. R. (2013). Invariant recurrent solutions embedded in a turbulent two-dimensional Kolmogorov flow. *Journal of Fluid Mechanics, 722*, 554–595. https://doi.org/10.1017/jfm.2013.122.

[7] Farazmand, M. (2016). An adjoint-based approach for finding invariant solutions of Navier Stokes equations. *Journal of Fluid Mechanics, 795*, 278–312. https://doi.org/10.1017/jfm.2016.203.

[8] Osborne, A. R. (2001). The random and deterministic dynamics of 'rogue waves' in unidirectional, deep-water wave trains. *Marine Structures, 14*, 275–293. https://doi.org/10.1016/S0951-8339(00)00064-2.

[9] Cousins, W., Sapsis, T. P. (2014). Quantification and prediction of extreme events in a one-dimensional nonlinear dispersive wave model. *Physica D: Nonlinear Phenomena, 280*, 48–58. https://doi.org/10.1016/j.physd.2014.04.012.

[10] Nieddu, G. T., Billings, L., Kaufman, J. H., Forgoston, E., Bianco, S. (2017). Extinction pathways and outbreak vulnerability in a stochastic Ebola model. *Journal of The Royal Society Interface, 14*, 20160847. https://doi.org/10.1098/rsif.2016.0847.

[11] Shaw, L. B., Schwartz, I. B. (2010). Enhanced vaccine control of epidemics in adaptive networks. *Physical Review E, 81*, 046120. https://doi.org/10.1103/PhysRevE.81.046120.

[12] Lloyd, A. L., May, R. M. (2001). How viruses spread among computers and people. *Science, 292*, 1316–1317. https://doi.org/10.1126/science.1061076.

[13] Belenky, V. L., Sevastianov, N. B. (2007). *Stability and safety of ships: Risk of capsizing*. The Society of Naval Architects and Marine Engineers.

[14] Mohamad, M. A., Sapsis, T. P. (2016). Probabilistic response and rare events in Mathieu's equation under correlated parametric excitation. *Ocean Engineering, 120*, 289–297. https://doi.org/10.1016/j.oceaneng.2016.03.008.

[15] Nayfeh, A. H., Mook, D. T. (1984). *Nonlinear oscillations*. Wiley-Interscience.

[16] Nicodemi, M. (2009). Extreme value statistics. In R. A. Meyers (Ed.), *Encyclopedia of complexity and systems science* (p. E:3317). *Springer*.

[17] Coles, S. G. (2001). An introduction to statistical modeling of extreme values (Springer Series in Statistics). *Springer*.

[18] Tit, K. (2024). Reliability of deep learning with rare event simulation: Theory and practice [Doctoral dissertation, Université de Rennes].

[19] Cui, T., Dolgov, S., Scheichl, R. (2024). Deep importance sampling using tensor trains with application to a priori and a posteriori rare events. SIAM Journal on Scientific Computing, 46(1). https://doi.org/10.1137/23M1546981.

[20] Dolgov, S., Anaya-Izquierdo, K., Fox, C., Scheichl, R. (2020). Approximation and sampling of multivariate probability distributions in the tensor train decomposition. *Statistics and Computing*, 30, 603–625.

[21] Cui, T., Dolgov, S. (2022). Deep composition of tensor-trains using squared inverse Rosenblatt transports. *Foundations of Computational Mathematics,* 22, 1863–1922.

[22] Bigoni, D., Engsign Karup, A. P., Marzouk, Y. M. (2016). Spectral tensor-train decomposition. *SIAM Journal on Scientific Computing*, 38, A2405–A2439.

[23] Gorodetsky, A., Karaman, S., Marzouk, Y. M. (2019). A continuous analogue of the tensor train decomposition. *Computer Methods in Applied Mechanics and Engineering,* 347, 59–84.

[24] Hackbusch, W. (2012). Tensor spaces and numerical tensor calculus (Vol. 42). *Springer Science & Business Media*.

[25] Oseledets, I., Tyrtyshnikov, E. (2010). TT-cross approximation for multidimensional arrays. *Linear Algebra and its Applications*, 432, 70–88.

[26] Oseledets, I. V. (2011). Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33, 2295–2317.

[27] Cui, T., Dolgov, S., Scheichl, R. (2022). Deep importance sampling using tensor trains with application to a priori and a posteriori rare event estimation. arXiv. https://arxiv.org/abs/2209.01941.