Breaking the curse of dimensionality in structured density estimation

NeurIPS 2024

Robert A. Vandermeulen, Wai Ming Tai, Bryon Aragam

TL;DR

- **Theoretically** how is it possible that some methods (e.g. neural networks) are so good at high dimensional data?
- Classic reasons:
  - Manifold hypothesis
  - Sparsity
    - Low-rank
  - Hierarchical assumptions
- This work: Conditional independence structure with graphical models (orthogonal to previous assumptions)
  - Effective dimension is related to a novel graph property

Nonparametric Density Estimation

- Problem: Given random vectors $X_1, X_2, \ldots, X_n \overset{iid}{\sim} p$
  - estimate p while making no (or very weak) assumptions on p
- If p is Lipschitz continuous the optimal rate for any estimator $\hat{p}$
  - $\dfrac{1}{\sqrt[d+2]{n}}$
- Can we improve this assuming a structured density?
  - Markov Random Field

Main Result

- Main result: if it is known that p satisfies the Markov property with respect to a graph G then there exists an estimator with the rate of $\frac{1}{\sqrt[2+r]{n}}$
  - $r(G)$ is a novel graph property we call "graph resilience"

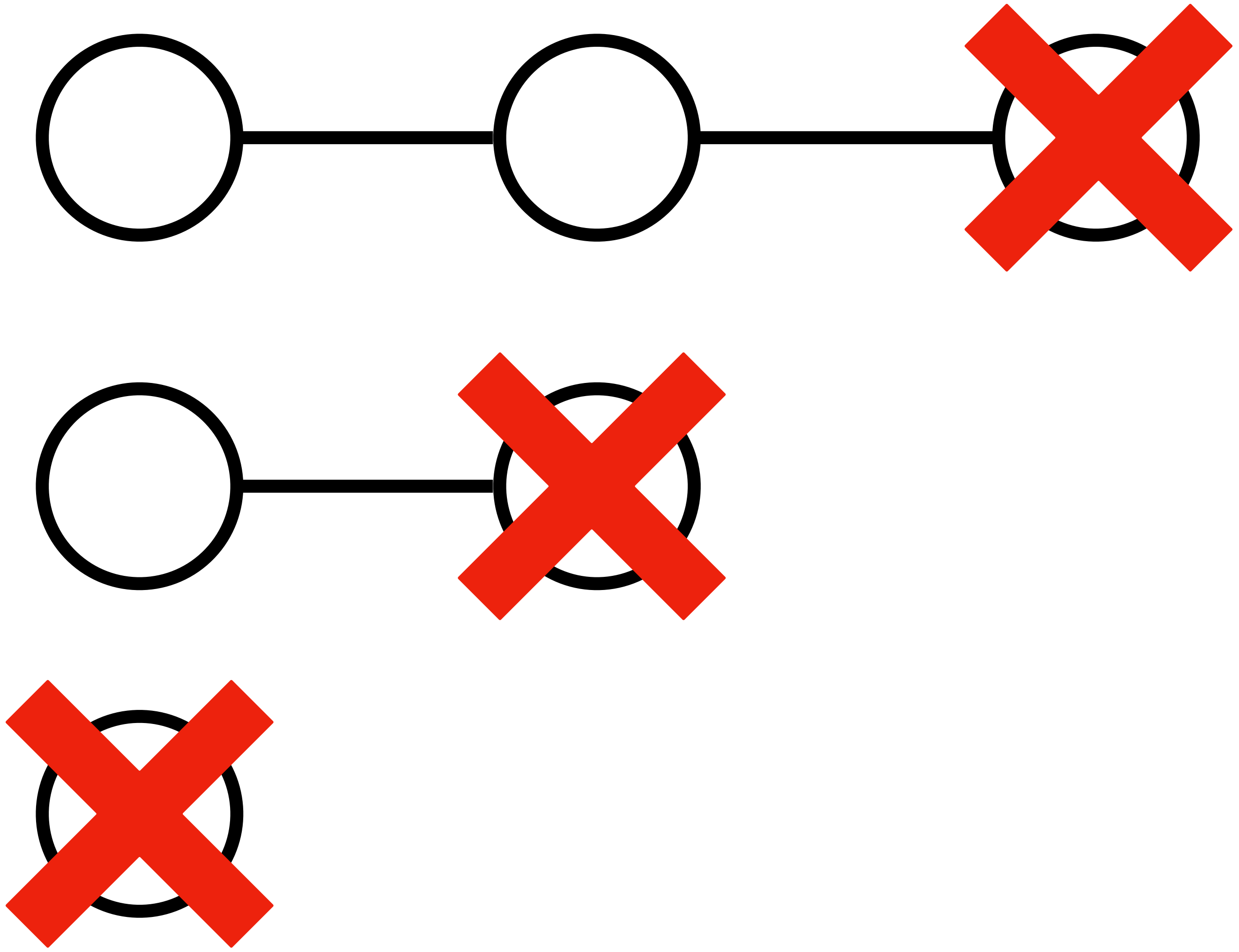
For nonparametric density estimation with a Markovian assumption G, there exists an estimator where the effective dimension is $r(G)$.

- $r(G) \leq d$
- For many reasonable G, $r(G) \ll d$

Graph Resilience: Definition

- Based on what we call a "graph disintegration"
- A disintegration of a graph G is a nested sequence of subgraphs $G \supsetneq G_1 \supsetneq G_2 \cdots \supsetneq G_\ell = \emptyset$
  - G_{i+1} has exactly one vertex removed from each component of G_i
  - ℓ is called the "length" of a disintegration
- There are typically many possible disintegrations of different possible lengths
- The resilience of a graph G is the length of its shortest disintegration

Disintegration Example



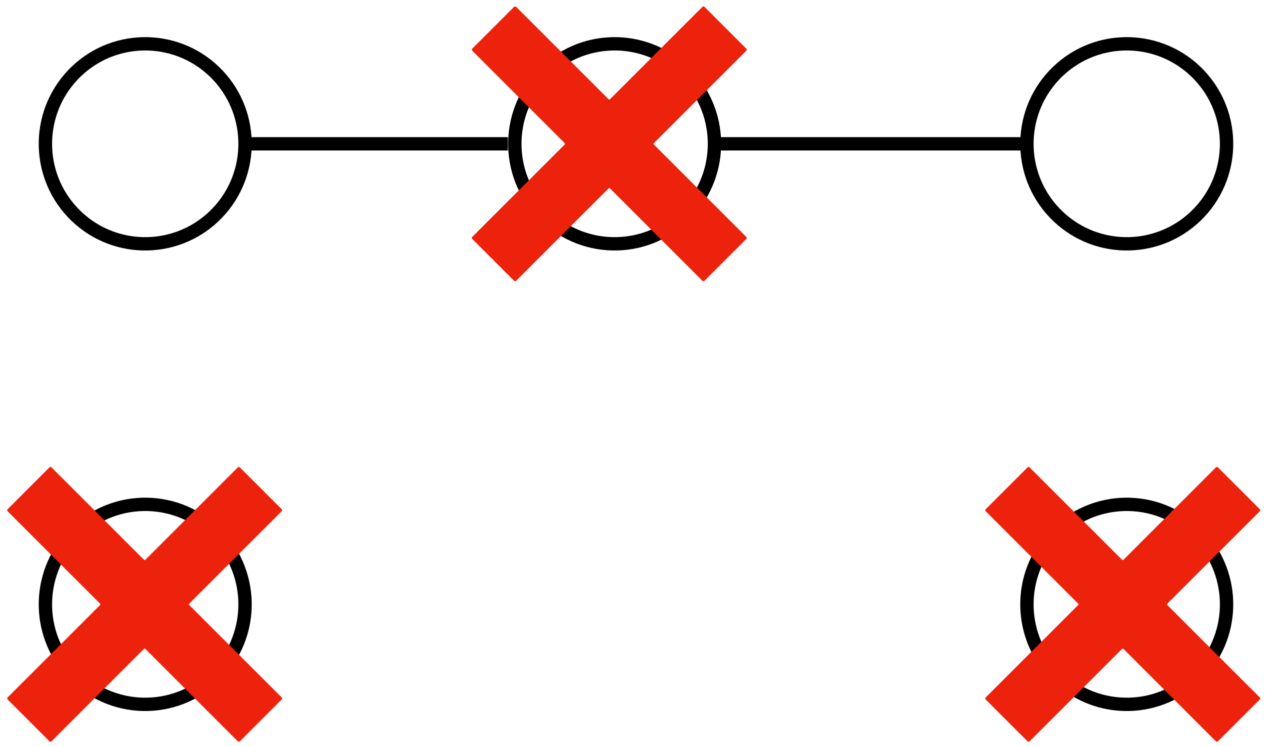
1

2

(Null graph)

3

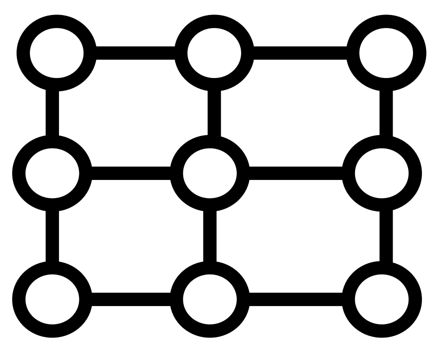
Disintegration Example



(Null graph)

2 steps in the shortest possible disintegration. Thus $r(G) = 2$.

Graph Resilience Bounds

- Path graph: 
    - $r(G) \leq \log_2(d) + 1$
- Complex sequential data, path graph where local dependence is expanded
    - $r(G) \lessapprox \log_2(d)$
- $k \times k$ Grid graph: 
    - $r(G) \leq \sqrt{d}$, where $d = k \cdot k$
- Complex spatial data: $r(G) \lessapprox \sqrt{d}$