Overview

1. Generic bounds on maximum deviations in sequential/sequence prediction
2. Viewpoint of “entropic innovations”
3. Implications in recursive algorithms and learning/generalization

- Entropy
- Information theory
- Innovations approach
- Estimation/prediction theory

Prediction Bound

Consider a stationary process \( \{ x_t \} , x_t \in \mathbb{R} \). Denote the 1-step ahead prediction of \( x_t \) by \( \hat{x}_t = f(\{ x_{t-k} \} | x_{t-1}) \). Then,

\[
D_{\text{max}}(x_t - \hat{x}_t) \leq \max_{i \in \text{supp}(x_t)} |x_t - \hat{x}_t| - E(x_t - \hat{x}_t)
\]

The maximum deviation:

\[
D_{\text{max}}(x_t - \hat{x}_t) \leq \max_{(x_{t-k}) \in \text{supp}(x_t)} |x_t - \hat{x}_t| - E(x_t - \hat{x}_t)
\]

For unbiased estimation:

\[
D_{\text{max}}(x_t - \hat{x}_t) = \max_{(x_{t-k}) \in \text{supp}(x_t)} |x_t - \hat{x}_t|
\]

Fundamental limitation of prediction: holds for arbitrary causal predictors

Viewpoint of “Entropic Innovations”

With \( \hat{x}_t = f(\{ x_{t-k} \} , x_{t-1}) \), it holds that

\[
l(x_t - \hat{x}_t; x_{t-1}) = l(x_t - \hat{x}_t; x_0, \ldots, x_{t-1} - \hat{x}_{t-1})
\]

Hence,

\[
l(x_t - \hat{x}_t; x_{t-1}) = 0
\]

\[
l(x_t - \hat{x}_t; x_{t-1}) = 0
\]

Prediction Bound for Stationary Processes

Consider a stationary process \( \{ x_t \} , x_t \in \mathbb{R} \). Denote the 1-step prediction of \( x_t \) by \( \hat{x}_t = f(\{ x_{t-k} \} | x_{t-1}) \). Then,

\[
\lim_{k \to \infty} \inf_{x_{t-k}} D_{\text{max}}(x_t - \hat{x}_t) \geq 2^{h_{\text{w}}(x_t)} - 1
\]

where equality holds iff \( x_t - \hat{x}_t \) is asymptotically white (strictly speaking, independent)

Perspective of entropic innovations:

\[
\lim_{k \to \infty} l(x_t - \hat{x}_t; x_{t-1}) = 0
\]

\[
\lim_{k \to \infty} l(x_t - \hat{x}_t; x_0, \ldots, x_{t-1} - \hat{x}_{t-1}) = 0
\]

\[
\{ x_t - \hat{x}_t \} \text{ is asymptotically white (strictly speaking, independent)}
\]

Optimal Predictor is “Uniformizing-Whitening”

\[
\lim_{k \to \infty} \inf_{x_{t-k}} D_{\text{max}}(x_t - \hat{x}_t) = 2^{h_{\text{w}}(x_t)} - 1
\]

holds iff the innovation process \( \{ x_t - \hat{x}_t \} \) is asymptotically white uniform.

- May feature an “uniformizing-whitening” principle

Implication 1: Recursive Algorithms

Consider a recursive algorithm given by

\[
x_{t+1} = x_t + f(x_{t-k}) + n_t
\]

where \( x_t \in \mathbb{R} \) denotes the recursive state, and \( n_t \in \mathbb{R} \) denotes the noise. Then,

\[
D_{\text{max}}(x_{t+1} - x_t) \geq 2^{h_{\text{w}}(y_{t-k})} - 1
\]

where equality holds iff \( x_{t+1} - x_t \) is uniform and \( l(x_{t+1} - x_t; n_{t-k}) = 0 \).

General Form

Consider a recursive algorithm given by

\[
g_{t+1}(x_{t-k+1}) + f(x_{t-k}) + n_t
\]

where \( x_t \in \mathbb{R} \) denotes the recursive state, and \( n_t \in \mathbb{R} \) denotes the noise. Then,

\[
D_{\text{max}}(g_{t+1}(x_{t-k+1}) + f(x_{t-k}) + n_t) \geq 2^{h_{\text{w}}(y_{t-k})} - 1
\]

where equality holds iff \( g_{t+1}(x_{t-k+1}) + f(x_{t-k}) + n_t \) is uniform and \( l(g_{t+1}(x_{t-k+1}) + f(x_{t-k}) + n_t; n_{t-k}) = 0 \).

First order:

\[
g_{t+1}(x_{t-k+1}) = x_{t+1} - x_t
\]

\[
x_{t+1} = x_t + f(x_{t-k}) + n_t
\]

Second order:

\[
x_{t+1} = x_{t+1} - x_t + f(x_{t-k}) + n_t
\]

Implication 2: Learning and Generalization

- Consider training data as input/output pairs \( (x_i, y_i), i = 0, \ldots, k \), where \( x_i \in \mathbb{R}^n \) is input and \( y_i \in \mathbb{R}^n \) is output
- Let the test input/output pair be \( (x_{test}, y_{test}) \), and denote the “prediction” (extrapolation/interpolation...) of \( y_{test} \) by \( \hat{y}_{test} = f(x_{test}) \), where \( f(\cdot) \) can be any learning algorithm
- Since the parameters of \( f(\cdot) \) are trained using \( (x_i, y_i), i = 0, \ldots, k \), eventually \( y_{test} \) is trained \( f(x_{test}) = g(x_{test}; y_{test}; x_{test}) \)

Then, for any learning algorithm \( f(\cdot) \),

\[
D_{\text{max}}(y_{test} - \hat{y}_{test}) \geq 2^{h_{\text{w}}(x_{test}; y_{test}; x_{test})} - 1
\]

where equality holds iff \( y_{test} = \hat{y}_{test} \) is uniform and \( l(y_{test} - \hat{y}_{test}; x_{test}, y_{test}; x_{test}) = 0 \).

Summary

- Fundamental limitations (generic bounds on maximum deviation) in prediction, recursive algorithms, and learning/generalization
- Future: How to achieve/approach?

References:

- T. M. Cover and J. A. Thomas
  Elements of Information Theory
- T. Kailath, A. H. Sayed, and B. Hassibi
  Linear Estimation
- P. P. Vaidyanathan
  The Theory of Linear Prediction
- S. Fang, J. Chen, and H. Ishi
  Towards Integrating Control and Information Theories: From Information-Theoretic Measures to Control Performance Limitations