Generic Bounds on the Maximum Deviations in Sequential/Sequence Prediction (and the Implications in Recursive Algorithms and Learning/Generalization)

Song Fang Quanyan Zhu

New York University

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 stochastic process $\{\mathbf{x}_k\}, \mathbf{x}_k \in \mathbb{R}$. Denote the 1-step ahead prediction of $\mathbf{x}_k$ by $\widehat{\mathbf{x}}_k = f_k(\mathbf{x}_{0,\ldots,k-1})$. Then,

$$\mathrm{D}_{\max}(\mathbf{x}_k - \widehat{\mathbf{x}}_k) \geq 2^{h(\mathbf{x}_k|\mathbf{x}_{0,\ldots,k-1})-1}$$

where equality holds iff $\mathbf{x}_k - \widehat{\mathbf{x}}_k$ is uniform and $I(\mathbf{x}_k - \widehat{\mathbf{x}}_k; \mathbf{x}_{0,\ldots,k-1}) = 0$.

- The maximum deviation:

$$\mathrm{D}_{\max}(\mathbf{x}_k - \widehat{\mathbf{x}}_k) \triangleq \max_{(\mathbf{x}_k - \widehat{\mathbf{x}}_k)\in \mathrm{supp}(\mathbf{x}_k - \widehat{\mathbf{x}}_k)} |\mathbf{x}_k - \widehat{\mathbf{x}}_k - \mathbb{E}(\mathbf{x}_k - \widehat{\mathbf{x}}_k)|$$

- For unbiased estimation:

$$\mathrm{D}_{\max}(\mathbf{x}_k - \widehat{\mathbf{x}}_k) = \max_{(\mathbf{x}_k - \widehat{\mathbf{x}}_k)\in \mathrm{supp}(\mathbf{x}_k - \widehat{\mathbf{x}}_k)} |\mathbf{x}_k - \widehat{\mathbf{x}}_k|$$

- Fundamental limitation of prediction; holds for arbitrary causal predictors

Viewpoint of "Entropic Innovations"

With $\widehat{\mathbf{x}}_k = f_k(\mathbf{x}_{0,\ldots,k-1})$, it holds that

$$I(\mathbf{x}_k - \widehat{\mathbf{x}}_k; \mathbf{x}_{0,\ldots,k-1}) = I(\mathbf{x}_k - \widehat{\mathbf{x}}_k; \mathbf{x}_0 - \widehat{\mathbf{x}}_0, \ldots, \mathbf{x}_{k-1} - \widehat{\mathbf{x}}_{k-1})$$

- Hence,

$$I(\mathbf{x}_k - \widehat{\mathbf{x}}_k; \mathbf{x}_{0,\ldots,k-1}) = 0$$
$$\Updownarrow$$
$$I(\mathbf{x}_k - \widehat{\mathbf{x}}_k; \mathbf{x}_0 - \widehat{\mathbf{x}}_0, \ldots, \mathbf{x}_{k-1} - \widehat{\mathbf{x}}_{k-1}) = 0$$

Prediction Bound for Stationary Processes

Consider a stationary process $\{\mathbf{x}_k\}, \mathbf{x}_k \in \mathbb{R}$. Denote the 1-step prediction of $\mathbf{x}_k$ by $\widehat{\mathbf{x}}_k = f_k(\mathbf{x}_{0,\ldots,k-1})$. Then,

$$\liminf_{k\to\infty} \mathrm{D}_{\max}(\mathbf{x}_k - \widehat{\mathbf{x}}_k) \geq 2^{h_\infty(\mathbf{x})-1}$$

where equality holds if $\{\mathbf{x}_k - \widehat{\mathbf{x}}_k\}$ is asymptotically uniform and $\lim_{k\to\infty} I(\mathbf{x}_k - \widehat{\mathbf{x}}_k; \mathbf{x}_{0,\ldots,k-1}) = 0$.

- Perspective of entropic innovations:

$$\lim_{k\to\infty} I(\mathbf{x}_k - \widehat{\mathbf{x}}_k; \mathbf{x}_{0,\ldots,k-1}) = 0$$
$$\Updownarrow$$
$$\lim_{k\to\infty} I(\mathbf{x}_k - \widehat{\mathbf{x}}_k; \mathbf{x}_0 - \widehat{\mathbf{x}}_0, \ldots, \mathbf{x}_{k-1} - \widehat{\mathbf{x}}_{k-1}) = 0$$
$$\Updownarrow$$
$$\{\mathbf{x}_k - \widehat{\mathbf{x}}_k\} \text{ is asymptotically white (strictly speaking, independent)}$$

Optimal Predictor is "Uniformizing-Whitening"

$$\liminf_{k\to\infty} \mathrm{D}_{\max}(\mathbf{x}_k - \widehat{\mathbf{x}}_k) = 2^{h_\infty(\mathbf{x})-1}$$

holds iff the innovation process $\{\mathbf{x}_k - \widehat{\mathbf{x}}_k\}$ is asymptotically white uniform.

- May feature an "uniformizing-whitening" principle

Implication 1: Recursive Algorithms

Consider a recursive algorithm given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k + f_k(\mathbf{x}_{0,\ldots,k}) + \mathbf{n}_k$$

where $\mathbf{x}_k \in \mathbb{R}$ denotes the recursive state, and $\mathbf{n}_k \in \mathbb{R}$ denotes the noise. Then,

$$\mathrm{D}_{\max}(\mathbf{x}_{k+1} - \mathbf{x}_k) \geq 2^{h(\mathbf{n}_k|\mathbf{n}_{0,\ldots,k-1})-1}$$

where equality holds iff $\mathbf{x}_{k+1} - \mathbf{x}_k$ is uniform and $I(\mathbf{x}_{k+1} - \mathbf{x}_k; \mathbf{n}_{0,\ldots,k-1}) = 0$.

General Form

Consider a recursive algorithm given by

$$g_{k+1}(\mathbf{x}_{0,\ldots,k+1}) = f_k(\mathbf{x}_{0,\ldots,k}) + \mathbf{n}_k$$

where $\mathbf{x}_k \in \mathbb{R}$ denotes the recursive state, and $\mathbf{n}_k \in \mathbb{R}$ denotes the noise. Then,

$$\mathrm{D}_{\max}[g_{k+1}(\mathbf{x}_{0,\ldots,k+1})] \geq 2^{h(\mathbf{n}_k|\mathbf{n}_{0,\ldots,k-1})-1}$$

where equality holds iff $g_{k+1}(\mathbf{x}_{0,\ldots,k+1})$ is uniform and $I(g_{k+1}(\mathbf{x}_{0,\ldots,k+1}); \mathbf{n}_{0,\ldots,k-1}) = 0$.

- First order:

$$g_{k+1}(\mathbf{x}_{0,\ldots,k+1}) = \mathbf{x}_{k+1} - \mathbf{x}_k$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k + f_k(\mathbf{x}_{0,\ldots,k}) + \mathbf{n}_k$$

- Second order:

$$g_{k+1}(\mathbf{x}_{0,\ldots,k+1}) = \mathbf{x}_{k+1} - 2\mathbf{x}_k + \mathbf{x}_{k-1}$$
$$\mathbf{x}_{k+1} = 2\mathbf{x}_k - \mathbf{x}_{k-1} + f_k(\mathbf{x}_{0,\ldots,k}) + \mathbf{n}_k$$
$$\vdots$$

Implication 2: Learning and Generalization

- Consider training data as input/output pairs $(\mathbf{x}_i, \mathbf{y}_i), i = 0, \ldots, k$, where $\mathbf{x}_i \in \mathbb{R}^n$ is input and $\mathbf{y}_i \in \mathbb{R}$ is output
- Let the test input/output pair be $(\mathbf{x}_{\text{test}}, \mathbf{y}_{\text{test}})$, and denote the "prediction" (extrapolation/interpolation...) of $\mathbf{y}_{\text{test}}$ by $\widehat{\mathbf{y}}_{\text{test}} = f(\mathbf{x}_{\text{test}})$, where $f(\cdot)$ can be any learning algorithm
- Since the parameters of $f(\cdot)$ are trained using $(\mathbf{x}_i, \mathbf{y}_i), i = 0, \ldots, k$, eventually $\widehat{\mathbf{y}}_{\text{test}} = f(\mathbf{x}_{\text{test}}) = g(\mathbf{x}_{\text{test}}, \mathbf{y}_{0,\ldots,k}, \mathbf{x}_{0,\ldots,k})$

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

$$\mathrm{D}_{\max}(\mathbf{y}_{\text{test}} - \widehat{\mathbf{y}}_{\text{test}}) \geq 2^{h(\mathbf{y}_{\text{test}}|\mathbf{x}_{\text{test}}, \mathbf{y}_{0,\ldots,k}, \mathbf{x}_{0,\ldots,k})-1}$$

where equality holds iff $\mathbf{y}_{\text{test}} - \widehat{\mathbf{y}}_{\text{test}}$ is uniform and $I(\mathbf{y}_{\text{test}} - \widehat{\mathbf{y}}_{\text{test}}; \mathbf{x}_{\text{test}}, \mathbf{y}_{0,\ldots,k}, \mathbf{x}_{0,\ldots,k}) = 0$.

Summary

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

- T. M. Cover and J. A. Thomas
Elements of Information Theory
John Wiley & Sons, 2006
- T. Kailath, A. H. Sayed, and B. Hassibi
Linear Estimation
Prentice Hall, 2000.
- P. P. Vaidyanathan
The Theory of Linear Prediction
Morgan & Claypool Publishers, 2007
- S. Fang, J. Chen, and H. Ishii
Towards Integrating Control and Information Theories: From Information-Theoretic Measures to Control Performance Limitations
Springer, 2017

song.fang@nyu.edu quanyan.zhu@nyu.edu