Statistical Contraction Theory for Learning-based Control of Nonlinear Systems—Part 3: Conformal Inference

Hiroyasu Tsukamoto

Department of Aerospace Engineering
University of Illinois Urbana-Champaign

February 3, 2026

Table of contents

- In the following, $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced 2-norm for matrices, i.e., $\|\cdot\| = \|\cdot\|_2$, unless otherwise noted.
- For a symmetric square matrix $S \in \mathbb{R}^{n\times n}$, we use the notation $S \succ 0$, $S \succeq 0$, $S \prec 0$, and $S \preceq 0$ for the positive definite, positive semi-definite, negative definite, and negative semi-definite matrices, respectively.
- This lecture is based on [1, 2, 3, 4].

Table of contents

Nonlinear Control with Formal Guarantees

What is missing?

Nonlinear Control with Formal Guarantees

Let's again consider the following perturbed nonlinear dynamical system:

$$\dot{\mathbf{x}} = f(\mathbf{x}, t) + d(\mathbf{x}, t)$$

where $d(\mathbf{x}, t)$ represents external disturbance with $\sup_{\mathbf{x},t} \|d(\mathbf{x}, t)\| = \bar{d} < \infty$.

Nonlinear Control with Formal Guarantees

What is unrealistic here?

Nonlinear Control with Formal Guarantees

How do we know that $d(\mathbf{x}, t)$ is bounded in the first place?

We could sample data of $d(\mathbf{x}, t)$ to empirically check if it is bounded, but do you think this is sufficient in safety-critical applications, e.g., in space?

Prediction Error Bound

This issue represents a central difficulty in machine learning.

Suppose a Llama large language model (LLM) outputs the following prompt:

```
================================================
Test prompt:  What should I do with my life?
Research topic:  aerospace robotics
================================================
Generating text before fine-tuning
================================================
What should I do with my life?  This is a question that has plagued many people
for years.  It's a question that can be both exciting and terrifying.  First,
let me tell you that there is no one-size-fits-all answer.  What works for one
person may not work for another.  What's most important is to take the time to
reflect on your values, passions, and strengths.  Ask yourself:
* What do I love doing?
* What makes me feel fulfilled?
* What ...
```

Prediction Error Bound

How can we verify that the LLM does not hallucinate, meaning its output is a valid and accurate response to the input prompt?

Prediction Error Bound

Suppose that we have a predictor for our system's uncertainty with N real-world data points describing the uncertainty.

Can we mathematically say anything about the uncertainty **without knowing anything about the prediction structure and data distribution**?

Table of contents

Statistical Inference

Definition (Statistic)

A statistic is any measurable function that depends only on the observed sampled data $(S_1, S_2, \cdots, S_N)$ *given as follows:*

$$y = T(S_1, S_2, \cdots, S_N).$$

For example,

$$T_{\text{ord}}(S_1, \cdots, S_N) = (S_{\pi(1)}, \cdots, S_{\pi(N)})$$

are referred to as the order statistics (where π is the rank permutation to be defined later).

Statistical Inference

The goal of statistical inference is to learn characteristics of a population from statistics of data samples.

This implies that statistical inference naturally encompasses all the (parametric) estimation approaches.

The term "estimation" typically refers to the process of determining unknown parameters or other quantities of interest that characterize the population.

Statistical Inference

The key difficulty in statistical inference is that **the characteristic parameters of the population are themselves unknown** because, at the beginning, everything is unknown.

Statistical Inference

Examples include, but are not limited to

- Probably approximately correct (PAC)-Bayes [5]
- Scenario approach [6, 7]
- Conformal inference [1, 2]

We focus on conformal inference, which may be most compatible with nonlinear control/estimation **without any structural and distributional priors**[1][2].

[1] T.-W. Hsu and H. Tsukamoto. "Statistical guarantees in data-driven nonlinear control: Conformal robustness for stability and safety". In: *IEEE Control Systems Letters* 9 (2025), pp. 997–1002. DOI: 10.1109/LCSYS.2025.3578062.

[2] S. Wei, M. Ornik, and H. Tsukamoto. "Conformal contraction for robust nonlinear control with distribution-free uncertainty quantification". In: *IEEE Conference on Decision and Control, to appear*. Dec. 2025. DOI: TBA.

Table of contents

Fundamentals of Order and Rank Statistics

Let $S_1, \cdots, S_N \in \mathbb{R}$ be real-valued random variables.

Definition (Exchangeability)

For any permutation $\sigma \in \mathcal{P}_N$,

$$(S_1, \cdots, S_N) \overset{d}{=} (S_{\sigma(1)}, \cdots, S_{\sigma(N)})$$

where $\mathcal{P}_N$ is the set containing all the permutations, $\overset{d}{=}$ means that the two sets of data samples have the same distribution. In particular, we have

$$\mathbb{P}[(S_1, \cdots, S_N) \in A] = \mathbb{P}[(S_{\sigma(1)}, \cdots, S_{\sigma(N)}) \in A]$$

for all [3] measurable sets A.

[3] the sets where the probability $\mathbb{P}[\cdot \in A]$ is well-defined

Fundamentals of Order and Rank Statistics

Definition (Rank Permutation)

Suppose that

- *The data samples* $(S_1, \cdots, S_N)$ *are exchangeable and*
- S_i *are almost surely distinct, i.e.,* $\mathbb{P}[S_i \neq S_j] = 1$ *for* $i \neq j$.

Let a_i *be the permuted indices with* $a_i \in \{1, \cdots, N\}$ *for each* $i \in \{1, \cdots, N\}$. *We define the rank permutation (a.k.a. ordering permutation) as*

$$\mathrm{Rank}(S_{a_1}, \cdots, S_{a_N}) = \pi \in \mathcal{P}_N$$

such that

$$S_{\pi(a_1)} < S_{\pi(a_2)} < \cdots < S_{\pi(a_N)}.$$

Fundamentals of Order and Rank Statistics

Definition (Rank Vector)

Suppose that

- *The data samples* $(S_1, \cdots, S_N)$ *are exchangeable and*
- S_i *are almost surely distinct, i.e.,* $\mathbb{P}[S_i \neq S_j] = 1$ *for* $i \neq j$.

We define the rank vector as

$$R = [R_1, \cdots, R_N]^\top, \text{ where } R_k = \#\, \ell \text{ s.t. } S_\ell < S_k \text{ for each } k = 1, \cdots, N.$$

Equivalently, this is given by

$$R = [R_1, \cdots, R_N]^\top, \text{ where } R_k = {\pi_{nom}}^{-1}(k) \text{ with } \pi_{nom} = \mathrm{Rank}(S_1, \cdots, S_N).$$

We call each R_k *the rank (or ranking) of the data sample* S_k.

Fundamentals of Order and Rank Statistics

Definition (Rank Statistics)

A statistic T of observed samples $(S_1, \cdots, S_N)$ defined only by their rank vector is called a rank statistic.

Fundamentals of Order and Rank Statistics

Theorem 1

Suppose that

- *The data samples* $(S_1, \cdots, S_N)$ *are exchangeable and*
- S_i *are almost surely distinct, i.e.,* $\mathbb{P}[S_i \neq S_j] = 1$ *for* $i \neq j$.

Then we have

$$\mathbb{P}[\mathrm{Rank}(S_{a_1}, \cdots, S_{a_N}) = \pi] = \frac{1}{N!} \text{ for all } \pi \in \mathcal{P}_N$$

where $a_i \in \{1, \cdots, N\}$ *for each* $i \in \{1, \cdots, N\}$ *denote arbitrary permuted indices.*

Fundamentals of Order and Rank Statistics

Proof: Let A_π be a set of $(S_{a_1}, \cdots, S_{a_N})$ defined as

$$A_\pi = \left\{ (S_{a_1}, \cdots, S_{a_N}) : S_{\pi(a_1)} < \cdots < S_{\pi(a_N)} \right\}$$

where $\pi \in \mathcal{P}_N$. Since we have for any permutation $\sigma \in \mathcal{P}_N$

$$(S_{a_1}, \cdots, S_{a_N}) \overset{d}{=} \left(S_{\sigma(a_1)}, \cdots, S_{\sigma(a_N)}\right)$$

due to exchangeability, we get

$$\mathbb{P}[(S_{a_1}, \cdots, S_{a_N}) \in A_\pi] = \mathbb{P}\left[\left(S_{\sigma(a_1)}, \cdots, S_{\sigma(a_N)}\right) \in A_\pi\right].$$

Fundamentals of Order and Rank Statistics

Also, by the definition of A_π, we have

$$\mathbb{P}[(S_{a_1}, \cdots, S_{a_N}) \in A_\pi] = \mathbb{P}\left[\text{Rank}\left(S_{a_1}, \cdots, S_{a_N}\right) = \pi\right]$$
$$\mathbb{P}\left[\left(S_{\sigma(a_1)}, \cdots, S_{\sigma(a_N)}\right) \in A_\pi\right] = \mathbb{P}\left[\text{Rank}\left(S_{\sigma(a_1)}, \cdots, S_{\sigma(a_N)}\right) = \pi\right]$$

and thus

$$\mathbb{P}\left[\text{Rank}\left(S_{a_1}, \cdots, S_{a_N}\right) = \pi\right] = \mathbb{P}\left[\text{Rank}\left(S_{\sigma(a_1)}, \cdots, S_{\sigma(a_N)}\right) = \pi\right].$$

Fundamentals of Order and Rank Statistics

Finally, for any $\pi, \varpi \in \mathcal{P}_N$, let us define σ as

$$\sigma = \pi^{-1} \circ \varpi \Rightarrow \varpi = \pi \circ \sigma.$$

Then we have

$$\begin{aligned}\text{Rank}\left(S_{\sigma(a_1)}, \cdots, S_{\sigma(a_N)}\right) = \pi &\Leftrightarrow S_{\pi(\sigma(a_1))} < \cdots < S_{\pi(\sigma(a_N))} \\ &\Leftrightarrow \text{Rank}\left(S_{a_1}, \cdots, S_{a_N}\right) = \varpi\end{aligned}$$

which yields

$$\mathbb{P}[\text{Rank}(S_{a_1}, \cdots, S_{a_N}) = \pi] = \mathbb{P}[\text{Rank}(S_{a_1}, \cdots, S_{a_N}) = \varpi]$$

for any $\pi, \varpi \in \mathcal{P}_N$. Using this relation with $|\mathcal{P}_N| = N!$ completes the proof. □

Fundamentals of Order and Rank Statistics

Corollary 2

Suppose again that

- *The data samples* $(S_1, \cdots, S_N)$ *are exchangeable and*
- S_i *are almost surely distinct, i.e.,* $\mathbb{P}[S_i \neq S_j] = 1$ *for* $i \neq j$.

Then we have

$$\mathbb{P}[R_k = r] = \frac{1}{N} \text{ for all } r = 1, \cdots, N$$

where R_k *is the* k*-th element of the rank vector.*

Proof: $|\mathcal{P}_N| = N!$ and we have $(N-1)!$ permutation options if we fix one of $(S_1, \cdots, S_N)$. □

Conformal Inference

Theorem 3 (Conformal Inference)

Suppose that

- *The data samples* $(S_1, \cdots, S_N, S_{N+1})$ *are exchangeable and*
- S_i *are almost surely distinct, i.e.,* $\mathbb{P}[S_i \neq S_j] = 1$ *for* $i \neq j$.

Let $\pi_{nom} = \mathrm{Rank}(S_1, \cdots, S_N)$. *Then we have*

$$\mathbb{P}\left[S_{N+1} \leq S_{\pi_{nom}(\lceil (1-\delta)(N+1) \rceil)}\right] \in \left[1-\delta, 1-\delta+\frac{1}{N+1}\right]$$

where $\delta \in (0,1)$ *and* $\lceil \cdot \rceil$ *is the ceiling function.*

Conformal Inference

Proof: We have

$$\mathbb{P}\left[S_{N+1} \leq S_{\pi_{nom}(\lceil(1-\delta)(N+1)\rceil)}\right] = \mathbb{P}\left[R_{N+1} \leq \lceil(1-\delta)(N+1)\rceil\right]$$
$$= \mathbb{P}\left[\bigcup_{r=1}^{\lceil(1-\delta)(N+1)\rceil} \{R_{N+1} = r\}\right] = \sum_{r=1}^{\lceil(1-\delta)(N+1)\rceil} \mathbb{P}\left[R_{N+1} = r\right] = \frac{\lceil(1-\delta)(N+1)\rceil}{N+1}$$

where R_{N+1} denotes the rank of S_{N+1}, the third equality is due to the fact that the events $\{R_{N+1} = r\}$ are disjoint, and the last equality is due to Corollary 2.

Conformal Inference

Then, by definition of the ceiling function, we have

$$\frac{\lceil (1-\delta)(N+1) \rceil}{N+1} \geq \frac{(1-\delta)(N+1)}{N+1} = 1-\delta$$

and

$$\frac{\lceil (1-\delta)(N+1) \rceil}{N+1} \leq \frac{(1-\delta)(N+1)+1}{N+1} = 1-\delta+\frac{1}{N+1}$$

which completes the proof. □

Conformal Inference

Note that the inequality

$$\mathbb{P}\left[S_{N+1} \leq S_{\pi_{nom}(\lceil (1-\delta)(N+1) \rceil)}\right] \in \left[1-\delta, 1-\delta+\frac{1}{N+1}\right]$$

can be interpreted as

$$\mathbb{P}\left[S_{N+1} \leq \lceil (1-\delta)(N+1) \rceil \text{-th smallest value in } (S_1, \cdots, S_N)\right] \in \left[1-\delta, 1-\delta+\frac{1}{N+1}\right].$$

Conformal Inference

Surprisingly, we did not assume anything about the underlying probability distribution. What's wrong with this bound?

Conformal Inference

The data you have $\{S_i\}_{i=1}^N$ ($= \{(x_i, y_i)\}_{i=1}^N$ for example) has to be exchangeable with the unknown, real-world data S_{N+1}.

The bound assumes randomness in $(S_1, \cdots, S_N)$, but in practice, this may be fixed ($\rightarrow$ beta distribution, see here or the appendix of this lecture note).

Table of contents

Examples

Consider an unknown function $\mathcal{F}$ with inputs $\mathcal{X}$ and outputs $\mathcal{Y}$

$$\mathcal{Y} = \mathcal{F}(\mathcal{X}) \tag{1}$$

and let $\hat{\mathcal{F}}$ be a prediction model of $\mathcal{F}$.

Examples

The function $\hat{\mathcal{F}}$ can represent various things.

- Machine learning models predicting real-world phenomena, e.g., stock prices
- Neural networks predicting uncertainty in dynamical systems
- Reinforcement learning agents predicting cumulative rewards
- LLMs/vision language models (VLMs) such as ChatGPT

Importantly, **we will not assume anything about the structure of $\mathcal{F}$ or the distribution of data points**.

Examples

Let us define the prediction nonconformity score as

$$s(\mathcal{X}, \mathcal{Y}) = \left\| \mathcal{Y} - \hat{\mathcal{F}}(\mathcal{X}) \right\|$$

where $\mathcal{Y}$ is the ground truth output given by (1).

Given data samples $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$, we quantify the prediction nonconformity scores for the conformal inference as

$$S_i = s(\mathcal{X}_i, \mathcal{Y}_i) = \left\| \mathcal{Y}_i - \hat{\mathcal{F}}(\mathcal{X}_i) \right\|, \ i = 1, \cdots, N.$$

Examples

Also, let

$$S_{N+1} = s(\mathcal{X}_{N+1}, \mathcal{Y}_{N+1}) = \left\| \mathcal{Y}_{N+1} - \hat{\mathcal{F}}(\mathcal{X}_{N+1}) \right\|$$

represent the nonconformity score of unseen, future data sample $(\mathcal{X}_{N+1}, \mathcal{Y}_{N+1})$.

Examples

Suppose that

- The data samples $(S_1, \cdots, S_N, S_{N+1})$ are exchangeable and
- The scores are almost surely distinct.

Let $\pi_{nom} = \mathrm{Rank}(S_1, \cdots, S_N)$. Then Theorem 3 implies that

$$\mathbb{P}\left[S_{N+1} \leq S_{\pi_{nom}(\lceil (1-\delta)(N+1) \rceil)}\right] \in \left[1-\delta, 1-\delta+\frac{1}{N+1}\right].$$

Examples

More intuitively, we have

$$\mathbb{P}\left[\left\|\mathcal{Y}_{N+1} - \hat{\mathcal{F}}(\mathcal{X}_{N+1})\right\| \leq \lceil (1-\delta)(N+1) \rceil \text{-th smallest in } (S_1, \cdots, S_N)\right] \in \left[1-\delta, 1-\delta+\frac{1}{N+1}\right].$$

We often just use the lower bound to get

$$\mathbb{P}\left[\left\|\mathcal{Y}_{N+1} - \hat{\mathcal{F}}(\mathcal{X}_{N+1})\right\| \leq \lceil (1-\delta)(N+1) \rceil \text{-th smallest in } (S_1, \cdots, S_N)\right] \geq 1-\delta.$$

Examples

In words, this means the unseen prediction error $\left\| \mathcal{Y}_{N+1} - \hat{\mathcal{F}}(\mathcal{X}_{N+1}) \right\|$ is bounded by $\lceil (1-\delta)(N+1) \rceil$-th smallest value in $(S_1, \cdots, S_N)$ at least with probability $1-\delta$.

This bound is no longer empirical but statistically rigorous. How can we use it with contraction theory?

Table of contents

Conditional Coverage

As mentioned earlier, **the bound in Theorem 3 assumes randomness in $(S_1, \cdots, S_N)$, but in practice, this may be fixed**.

Let's briefly see what happens if this is the case, using the conditional cumulative distribution function (CDF) of S_{N+1} defined as

$$F_S(x) = \mathbb{P}[S_{N+1} \leq x \mid S_1, \cdots, S_N].$$

Conditional Coverage

Theorem 4 (Conditional Coverage)

Suppose that

- *The data samples* $(S_1, \cdots, S_N, S_{N+1})$ *are i.i.d. and*
- *The common CDF of* $S_i,\ i = 1, \cdots, N+1$ *is strictly increasing and continuous.*

Then the following random variable

$$F_S\left(S_{\pi_{nom}(k)}\right) = \mathbb{P}\left[S_{N+1} \leq S_{\pi_{nom}(k)} \mid S_1, \cdots, S_N\right]$$

is beta distributed. Note that the probability itself is a random variable here.

Conditional Coverage

Proof: Since the common CDF is continuous and strictly increasing, we have

$$\{F_S(S_{\pi_{nom}(k)}) \leq p\} \Leftrightarrow \{S_{\pi_{nom}(k)} \leq s\}$$

for $p \in (0,1)$ and $s = F_S^{-1}(p)$, where $k \in \{1, \cdots, N\}$. More intuitively, we have

$$\{F_S(S_{\pi_{nom}(k)}) \leq p\} \Leftrightarrow \{S_{\pi_{nom}(k)} \leq s\} \Leftrightarrow \{\underline{\text{at least}}\ k \text{ of the } N \text{ samples are} \leq s\}$$

$$\Leftrightarrow \left\{ \sum_{i=1}^{N} \mathbb{1}\{S_i \leq s\} \geq k \right\}$$

where $\mathbb{1}$ is the indicator function.

Conditional Coverage

Since the data samples are i.i.d., we have $\mathbb{P}[S_i \leq s] = F_S(s) = p$ and then $\mathbb{P}[S_i > s] = 1 - p$, which implies $\mathbb{1}\{S_i \leq s\} \sim$ Bernoulli(p) for each $i = 1, \cdots, N$.

This further gives $\sum_{i=1}^{N} \mathbb{1}\{S_i \leq s\} \sim$ Binomial(N, p), i.e.,

$$\mathbb{P}\left[\sum_{i=1}^{N} \mathbb{1}\{S_i \leq s\} \geq k\right] = \sum_{j=k}^{N} \binom{N}{j} p^j (1-p)^{N-j}.$$

Conditional Coverage

Thus, we get

$$\mathbb{P}\left[F_S(S_{\pi_{nom}(k)}) \leq p\right] = \sum_{j=k}^{N} \binom{N}{j} p^j (1-p)^{N-j}.$$

Viewing $P = F_S(S_{\pi_{nom}(k)})$ as a random variable, we can also see that $F_P(p) = \mathbb{P}\left[F(S_{\pi_{nom}(k)}) \leq p\right] = \mathbb{P}[P \leq p]$ is the CDF of P.

Therefore, the probability density function (PDF) of P is

$$\frac{dF_P}{dp} = \frac{N!}{(k-1)!(N-k)!} p^{k-1}(1-p)^{N-k}.$$

Conditional Coverage

Finally, let $\alpha = k$ and $\beta = N - k + 1$. For positive integers n, the Gamma function satisfies $\Gamma(n) = (n-1)!$, which gives

$$\frac{dF_P}{dp} = \frac{\Gamma(N+1)}{\Gamma(k)\Gamma(N-k+1)} p^{k-1}(1-p)^{N-k} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

which is indeed the PDF of the beta distribution. □

Reference

[1] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer US, 2005. DOI: 10.1007/978-0-387-22732-3.

[2] R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. "Conformal prediction under covariate shift". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates Inc., 2019.

[3] T.-W. Hsu and H. Tsukamoto. "Statistical guarantees in data-driven nonlinear control: Conformal robustness for stability and safety". In: *IEEE Control Systems Letters* 9 (2025), pp. 997–1002. DOI: 10.1109/LCSYS.2025.3578062.

[4] S. Wei, M. Ornik, and H. Tsukamoto. "Conformal contraction for robust nonlinear control with distribution-free uncertainty quantification". In: *IEEE Conference on Decision and Control, to appear*. Dec. 2025. DOI: TBA.

Reference

[5] D. A. McAllester. "Some PAC-Bayesian theorems". In: *Machine Learning* 37.3 (1999), pp. 355–363. ISSN: 1573-0565. DOI: 10.1023/A:1007618624809.

[6] G. C. Calafiore and M. C. Campi. "The scenario approach to robust control design". In: *IEEE Transactions on Automatic Control* 51.5 (2006), pp. 742–753. DOI: 10.1109/TAC.2006.875041.

[7] M. C. Campi, S. Garatti, and M. Prandini. "The scenario approach for systems and control design". In: *Annual Reviews in Control* 33.2 (2009), pp. 149–157. DOI: 10.1016/j.arcontrol.2009.07.001.