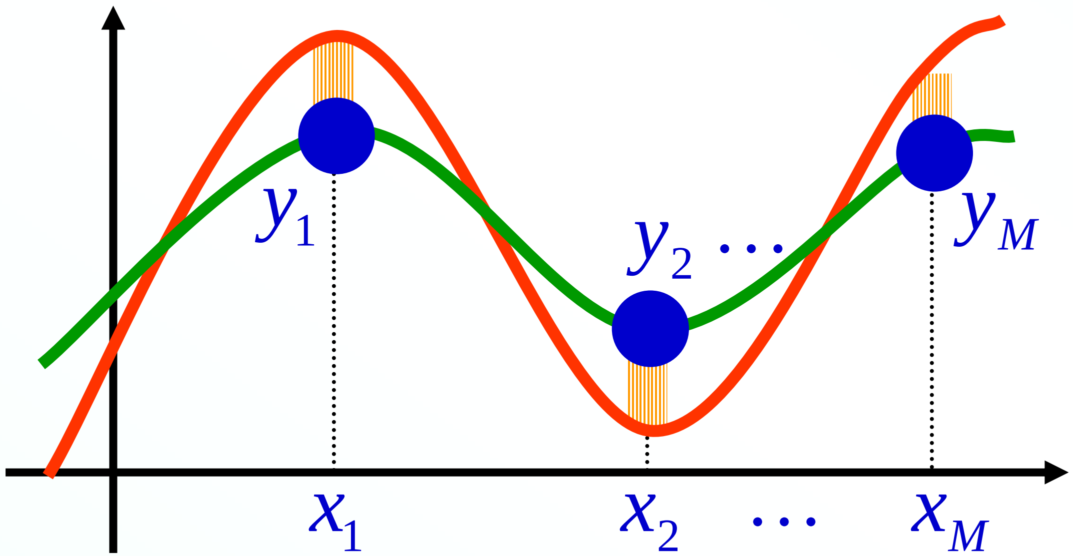
IJCNN2002 May 12-17, 2002

Release from Active Learning / Model Selection Dilemma: Optimizing Sample Points and Models at the Same Time

Department of Computer Science,
Tokyo Institute of Technology, Tokyo, Japan

Masashi Sugiyama
Hidemitsu Ogawa

Supervised Learning: Function Approximation



$f(x)$: Learning target

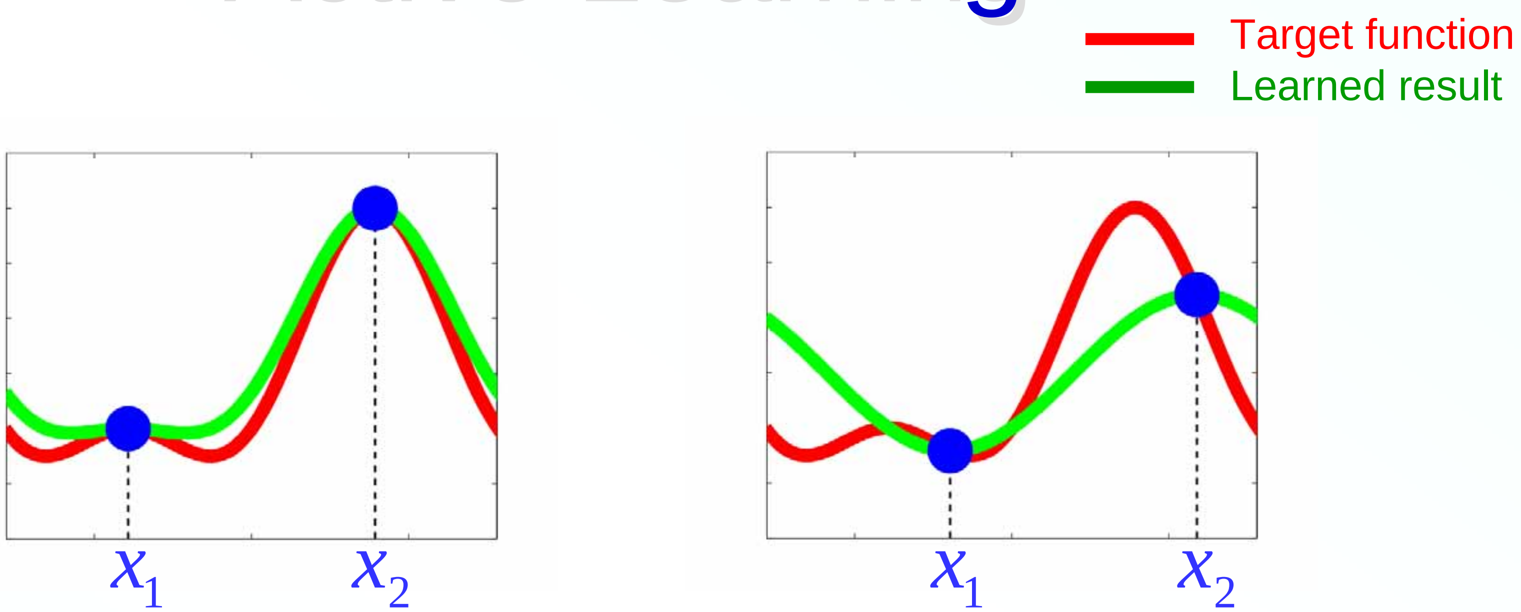
$\hat{f}(x)$: Learned result

$\{x_m, y_m\}_{m=1}^{M}$: Samples

$$y_m = f(x_m) + \varepsilon_m$$

From $\{x_m, y_m\}_{m=1}^{M}$, find $\hat{f}(x)$
so that it is as close to $f(x)$ as possible

Active Learning



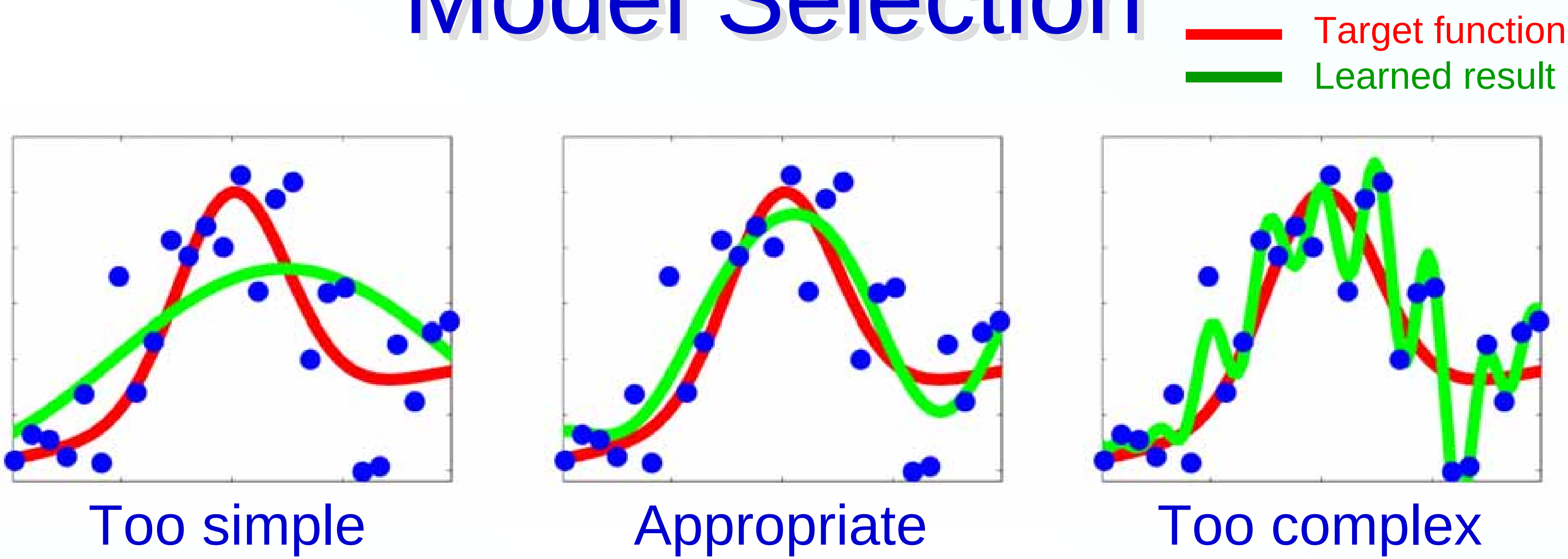
Location of sample points AFFECTS heavily

Determine $\{x_m\}_{m=1}^{M}$ for optimal generalization

$$\min_{\{x_m\}} J_G$$

J_G : Generalization error

Model Selection



Choice of models AFFECTS heavily

(Model refers to, e.g., order of polynomials)

Select a model S for optimal generalization

$$\min_{S \in C} J_G$$

C : Set of model candidates

J_G : Generalization error

Simultaneous Optimization of Sample Points and Models

So far, active learning and model selection have been studied thoroughly, but INDEPENDENTLY

Simultaneously determine sample points $\{x_m\}_{m=1}^{M}$ and a model S for optimal generalization

$$\min_{\{x_m\},\ S\in C} J_G$$

C : Set of model candidates

J_G : Generalization error

Active Learning / Model Selection Dilemma

We can NOT directly optimize
sample points and models simultaneously
by simply combining existing
active learning and model selection methods

Because…

- Model should be fixed for active learning
- Sample points should be fixed for model selection

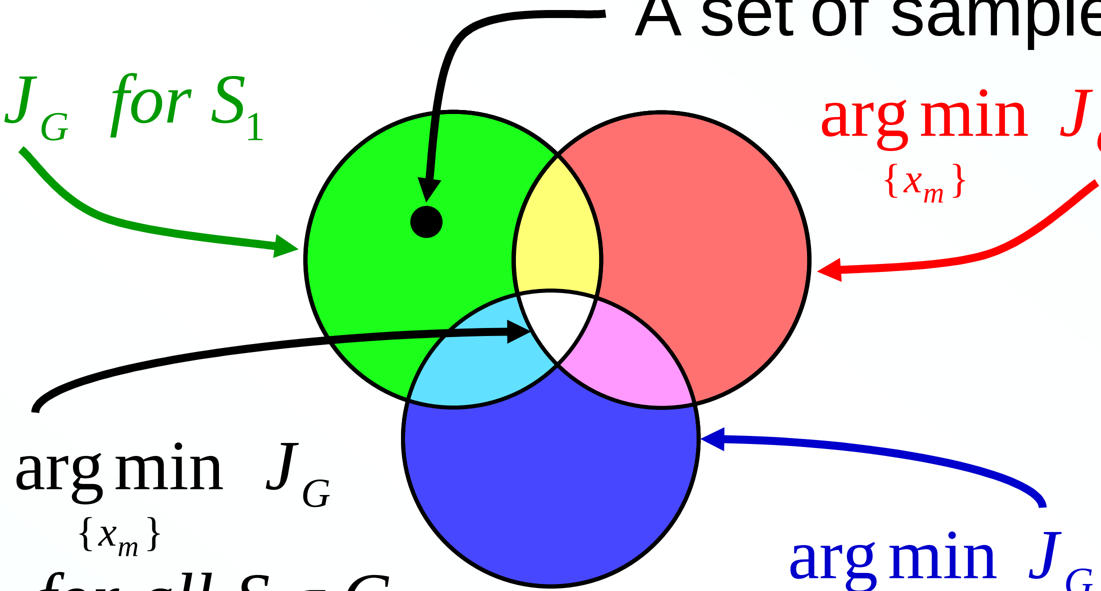
How to Dissolve the Dilemma

Model candidates : $C = \{S_1, S_2, S_3\}$

A set of sample points $\{x_m\}$

$\underset{\{x_m\}}{\arg\min}\ J_G\ \ for\ S_1$

$\underset{\{x_m\}}{\arg\min}\ J_G\ \ for\ S_2$



$\{x_m^{\text{(OPT)}}\} = \underset{\{x_m\}}{\arg\min}\ J_G$

$for\ all\ S \in C$

$\underset{\{x_m\}}{\arg\min}\ J_G\ \ for\ S_3$

1. Find sample points $\left\{x_m^{(OPT)}\right\}_{m=1}^{M}$ that are commonly optimal for all models
2. Just perform model selection as usual

Is It Just Idealistic?

No! Commonly optimal sample points surely exist for trigonometric polynomial models

Trigonometric polynomial model of order n

$$\hat{f}(x) = \theta_1 + \sum_{p=1}^{n} \left(\theta_{2p} \sin px + \theta_{2p+1} \cos px \right)$$

From here on, we assume

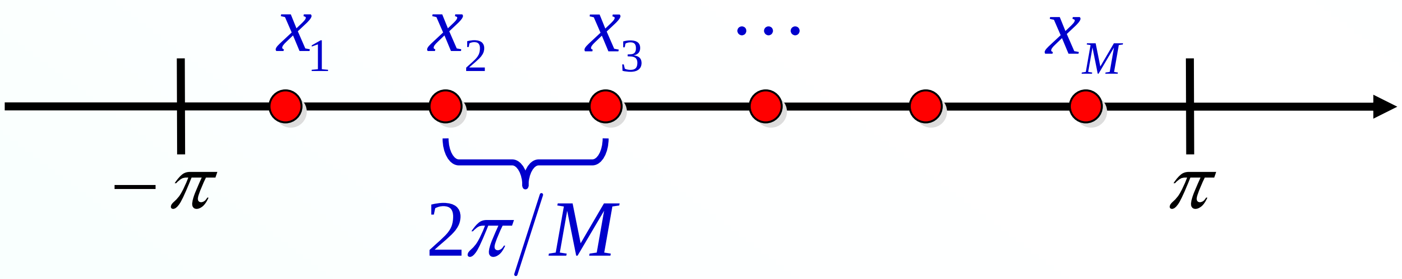
- Least mean squares (LMS) estimate
- Generalization measure: $J_G = E \int_{-\pi}^{\pi} \left| \hat{f}(x) - f(x) \right| dx$

E : Expectation over noise

Theorem

For all trigonometric polynomial models that include learning target function, equidistance sampling gives the optimal generalization capability

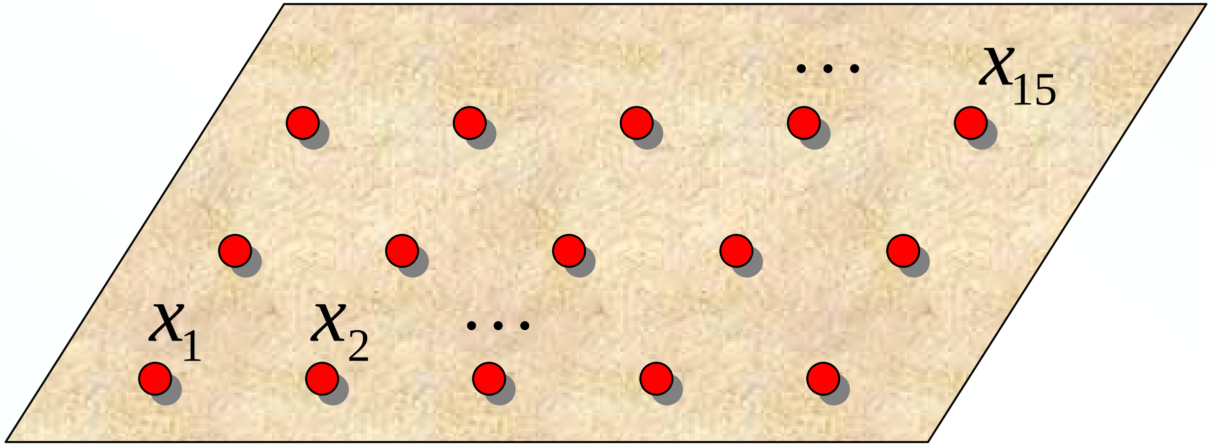
1-dimensional input



M : Number of samples

Multi-Dimensional Input Cases

2-dimensional input



Sampling on regular grid is optimal

Computer Simulations (Artificial, Realizable)

- Learning target function: $f \in S_{50}$

  S_n : Trigonometric polynomial model of order n

- Model candidates: $C = \{S_0, S_1, S_2, \ldots, S_{100}\}$

- Generalization measure:

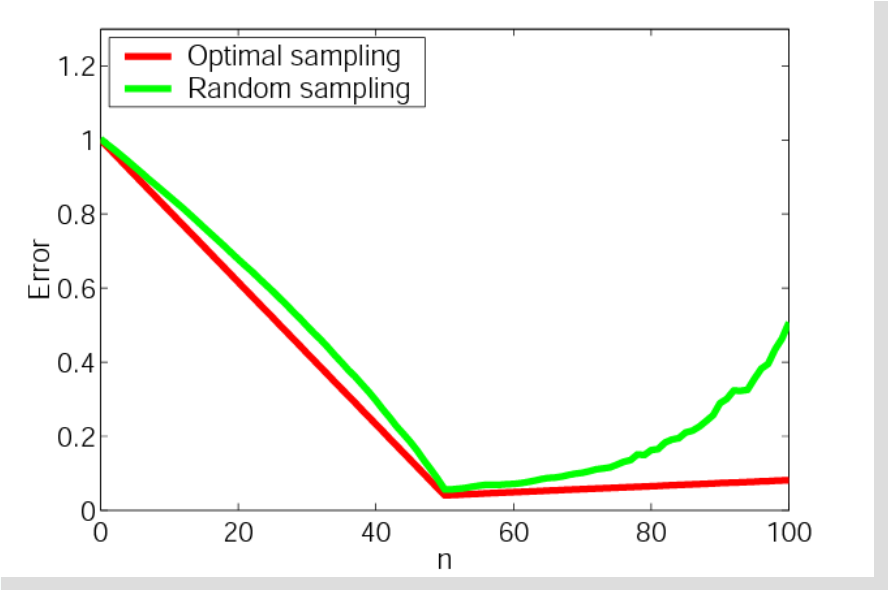
$$J_G = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \hat{f}(x) - f(x) \right| dx$$

- Sampling schemes:
  - Equidistance sampling
  - Random sampling

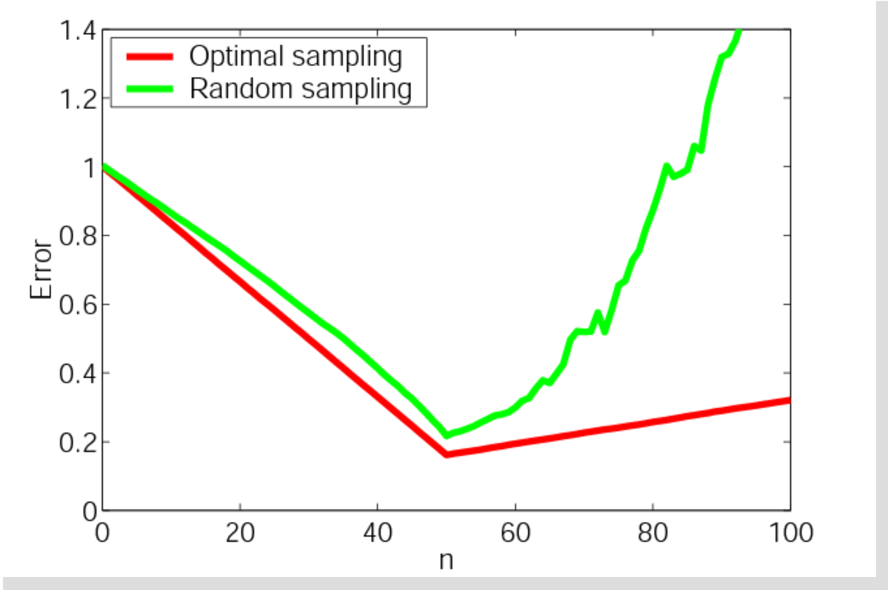
Simulation Results (Large Samples)

Number of samples $= 500$

Noise variance $= 0.02$



Noise variance $= 0.08$



Horizontal: Order of models

Vertical: Generalization error

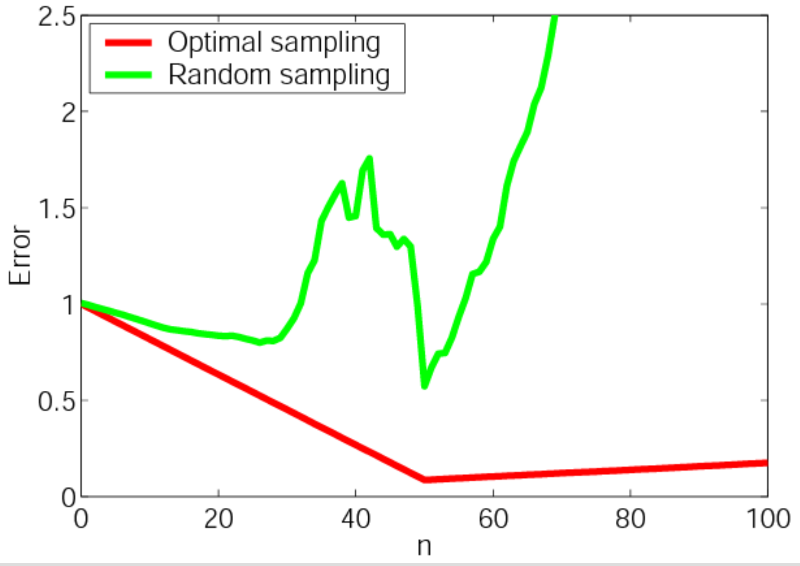
Averaged over 100 trials

Equidistance sampling outperforms random sampling for all models!

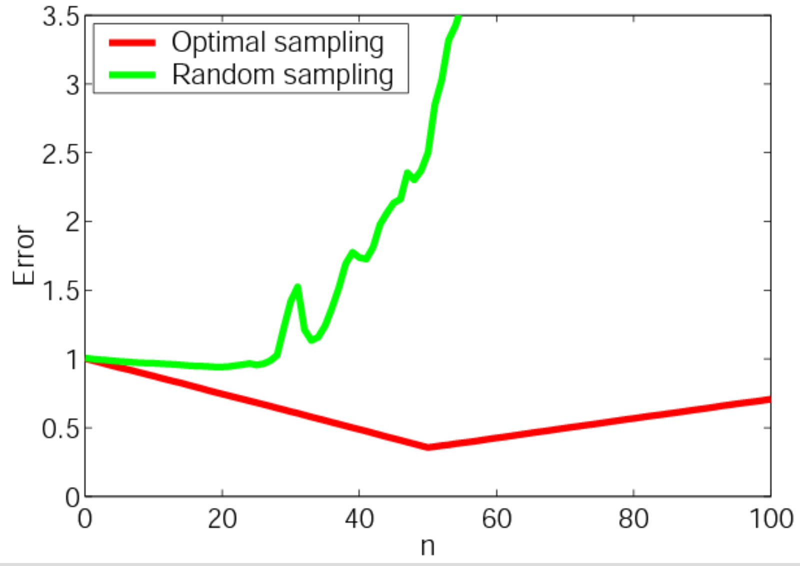
Simulation Results (Small Samples)

Number of samples $= 230$

Noise variance $= 0.02$



Noise variance $= 0.08$



Horizontal: Order of models

Vertical: Generalization error

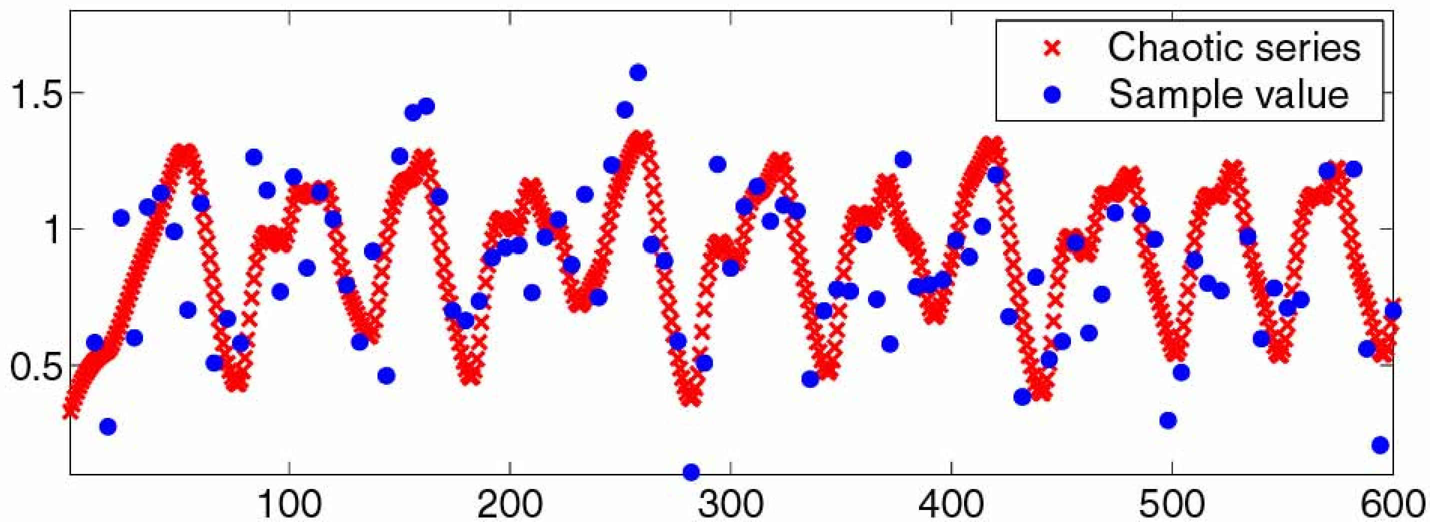
Averaged over 100 trials

With small samples, equidistance sampling performs excellently for all models!

Computer Simulations (Unrealizable)

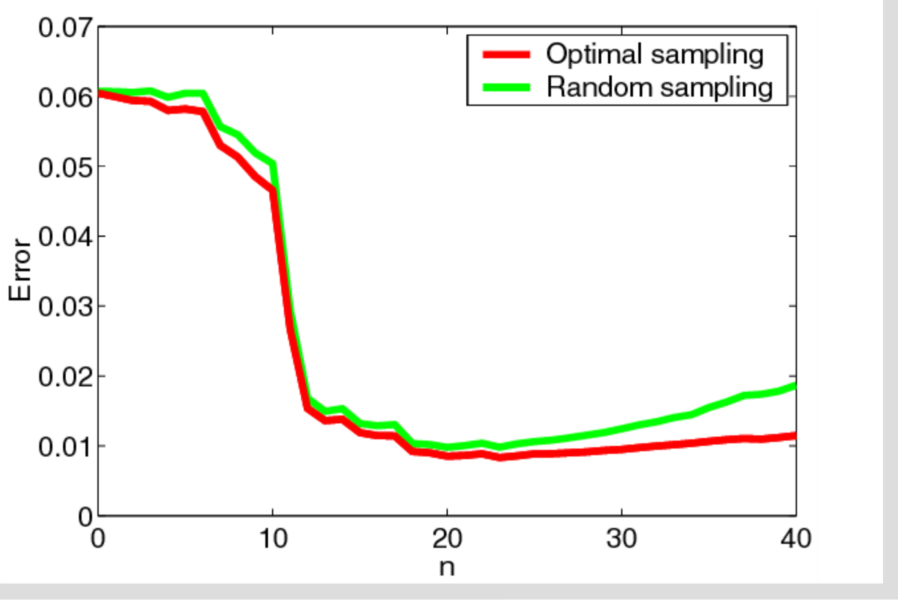
- Interpolate 600 chaotic series (red) from noisy samples (blue)

- Model candidates: $C = \{S_0, S_1, S_2, \ldots, S_{40}\}$
  S_n : Trigonometric polynomial model of order n

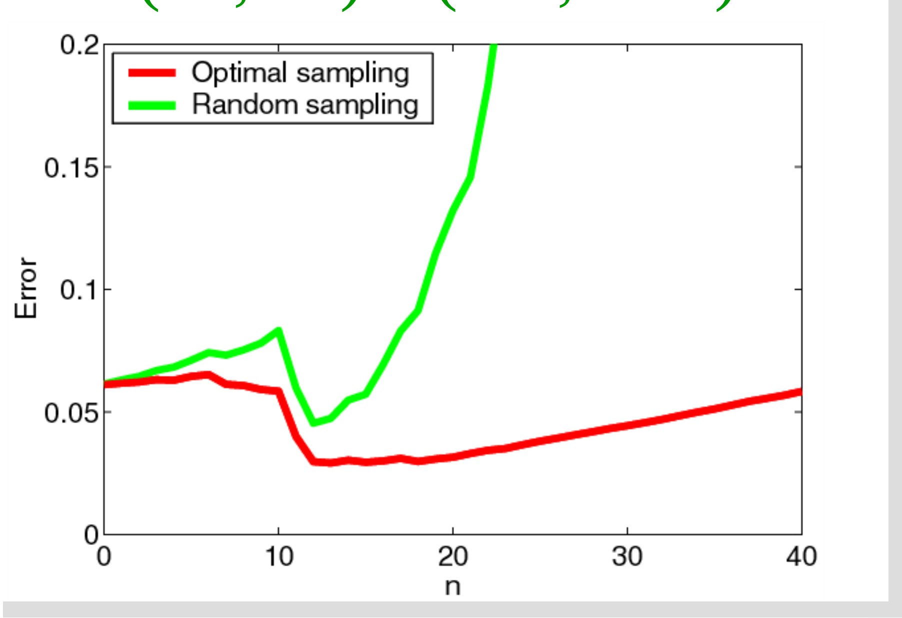


Simulation Results (Unrealizable)

$(M, \sigma^2) = (300, 0.04)$



$(M, \sigma^2) = (100, 0.07)$



Horizontal: Order of models

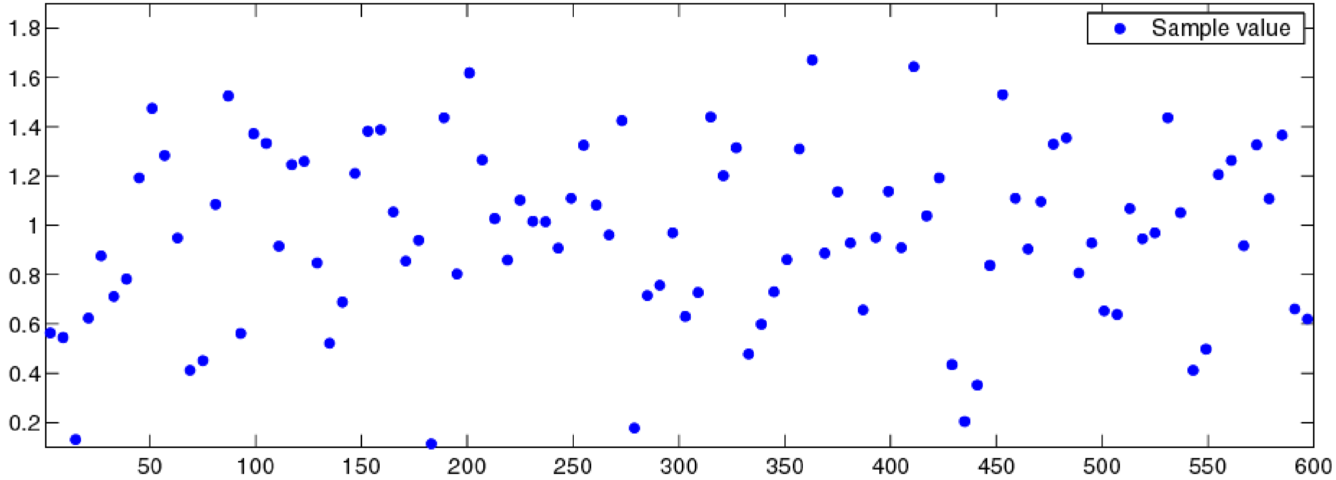
Vertical: Test error at all 600 points

Averaged over 100 trials

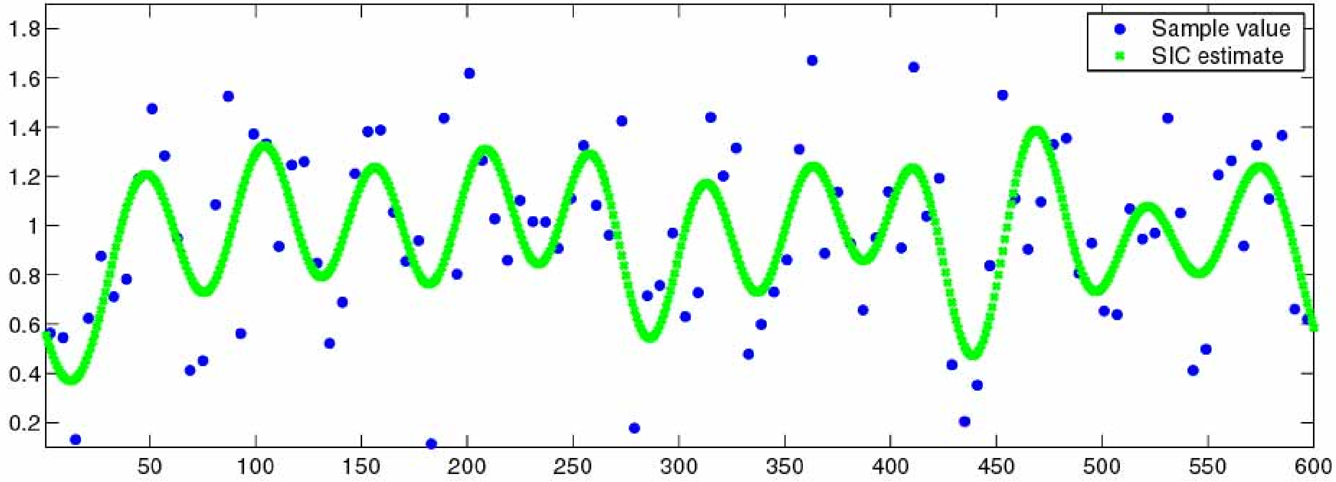
Equidistance sampling outperforms random sampling for all models!

Interpolated Chaotic Series

After model selection with equidistance sampling,

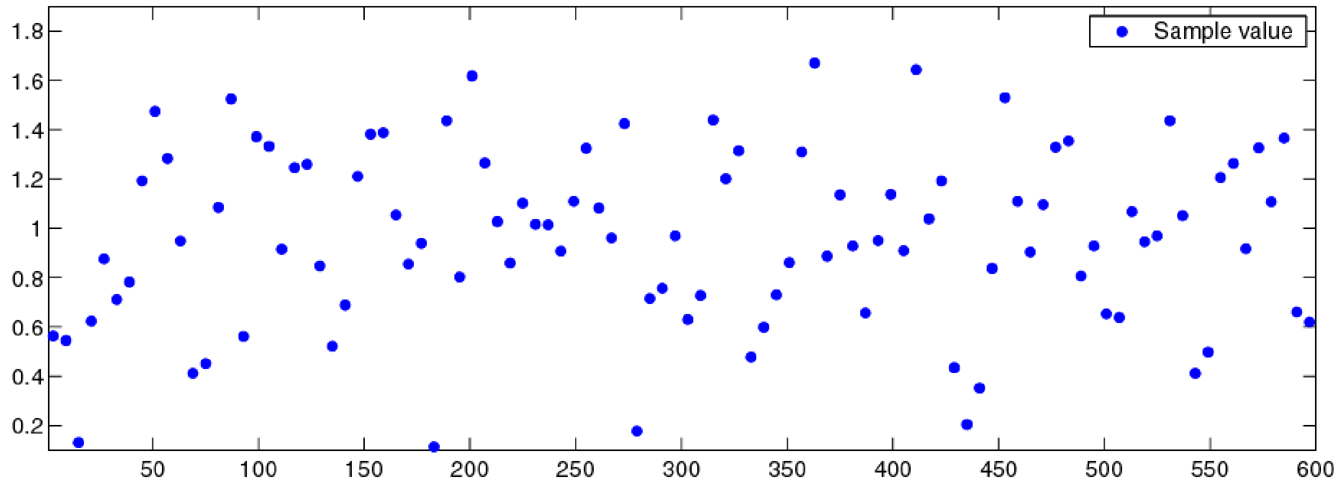


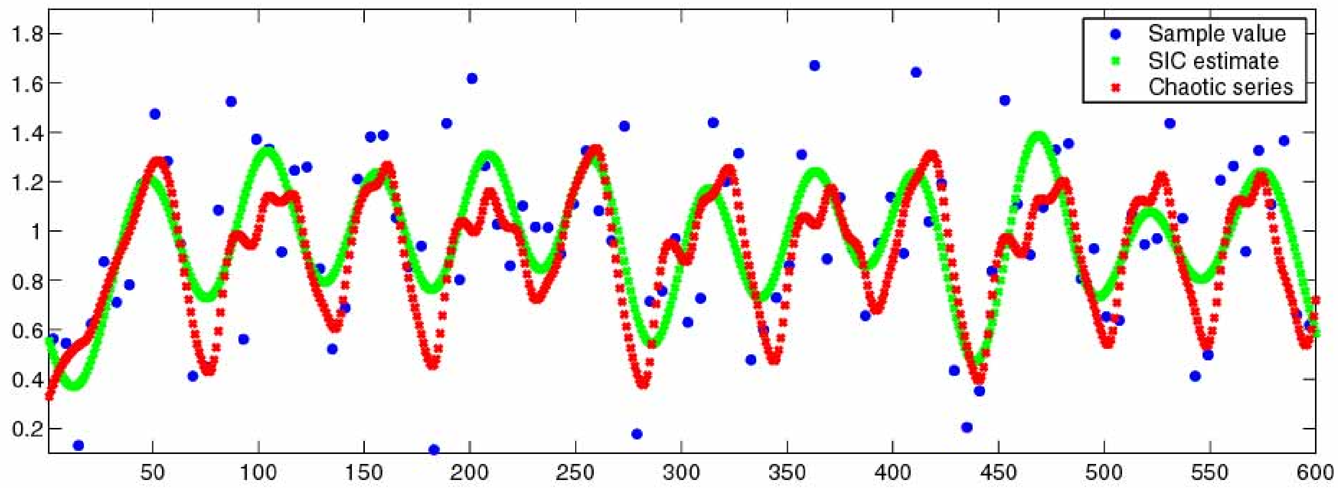
Selected model : S_{13}



Compared with True Series

We obtained good estimates from sparse data!





Conclusions

- **Active learning / model selection dilemma:**
  Sample points and models can not be simultaneously optimized by simply combining existing active learning and model selection methods

- **How to dissolve the dilemma:**
  Find commonly optimal sample points for all models

- **Is it realistic?**
  Commonly optimal sample points surely exist for trigonometric polynomial models: equidistance sampling

- **Is it practical?**
  Computer simulations showed that the proposed method works excellently even in unrealizable cases