

Sequential Model List Selection for Function Approximation

Ernest Fokoué

epf@samsi.info



Joint work with

Bertrand Clarke

UBC, SAMSI, Duke

Outline of the Presentation

- General Introduction

Outline of the Presentation

- General Introduction
- Sources of Uncertainty

Outline of the Presentation

- General Introduction
- Sources of Uncertainty
- Appeal of Model Averaging

Outline of the Presentation

- General Introduction
- Sources of Uncertainty
- Appeal of Model Averaging
- Pitfalls of Naive Averages

Outline of the Presentation

- General Introduction
- Sources of Uncertainty
- Appeal of Model Averaging
- Pitfalls of Naive Averages
- A Sequential Selection Solution

Outline of the Presentation

- General Introduction
- Sources of Uncertainty
- Appeal of Model Averaging
- Pitfalls of Naive Averages
- A Sequential Selection Solution
- Illustrative Examples

Outline of the Presentation

- General Introduction
- Sources of Uncertainty
- Appeal of Model Averaging
- Pitfalls of Naive Averages
- A Sequential Selection Solution
- Illustrative Examples
- Conclusion and Future Work

General Problem Formulation

- Given iid data $D = \{(\mathbf{x}_i, y_i)_{i=1}^n\}$ where

$$Y_i = f^*(\mathbf{x}_i) + \epsilon_i$$

- Function approximation

$$\text{Find } f^{\text{opt}} = \arg \min_{f \in \mathcal{F}} R(f)$$



- $R(f) = \mathbb{E}_{\mathbf{x}y} [(Y - f(X))^2]$ is risk functional.

- Prediction error $R(f)$

$$R(f) = \frac{1}{m} \sum_{i=1}^m (y_i^{\text{new}} - f(\mathbf{x}_i^{\text{new}}))^2$$

- How to find this predictively optimal function f ?

Basis Expansion Approach

- Basis function set

$$\mathcal{E} = \{e_1, e_2, \dots, e_k\}$$

- Function space

$$\mathcal{F} \equiv \text{span } \mathcal{E}$$

- $\forall f \in \mathcal{F}, \exists p \in \{1, 2, \dots, k\}$

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^p \beta_j e_j(\mathbf{x}) \quad (1)$$

- Model space:

$$\mathbb{M} = \{M : \text{where } M \text{ models a function of the form (1)}\}$$

Sources of Uncertainty

- Parameter uncertainty
 - For any given model $M \in \mathbb{M}$, there is uncertainty in its parameters. Bayesian inference takes care of this uncertainty very well?
- Model uncertainty
 - Given a list $\mathcal{M} \subset \mathbb{M}$ of "plausible" models from \mathbb{M} , different models will produce different predictions. Model Averaging and Model Selection help account for model uncertainty?
- Model list uncertainty
 - For a class of models in a model space \mathbb{M} , how do we select a list \mathcal{M} of plausible models? Topic always ignored!

The Appeal of Model Averaging

Bayesian Model Averaging is well established as the optimal predictive solution in function approximation

So, if predictive optimality is the **will**, then Bayesian Model Averaging would seem to be the **way**

Pitfalls of Naive Model Averaging

- It happens that from the same model space \mathbb{M} some model lists produce higher prediction errors than others ...
- Careless prior specification on a single model list can denigrate the model average obtained from it.
- Arbitrary large model lists have been seen to increase the average prediction error.

Note: Model list variability has not been given the proper care that it deserves.

Note: This work argues that a selective model averaging might be the way to negotiate a bias-variance trade-off so as to drive the prediction error as small as possible.

Pitfalls of Naive Averages (I)

Existence of regions of high redundancy in model space

- **Cause:** Highly correlated predictors or linearly dependent basis functions.
- **Consequence:** Uniformity of $p(M)$ leads to skewness of $p(M|D)$: **Averages suspicious.**
- **A remedy:** Dilution priors by Ed George

Dilution Priors

Assign prior probabilities uniformly to model neighborhoods.

- Bayesian Linear Model
 - Voronoi tessellation of full model space.
- Bayesian CART
 - Tree-generating process priors (CART).

Note:

- Such priors do not require subjective inputs.

Pitfalls of Naive Averages (II)

Vague convergence to zero

- Causes:
 - Model list far larger than n
 - Uniform prior $p(M)$
 - Large list of similar models

- Consequence

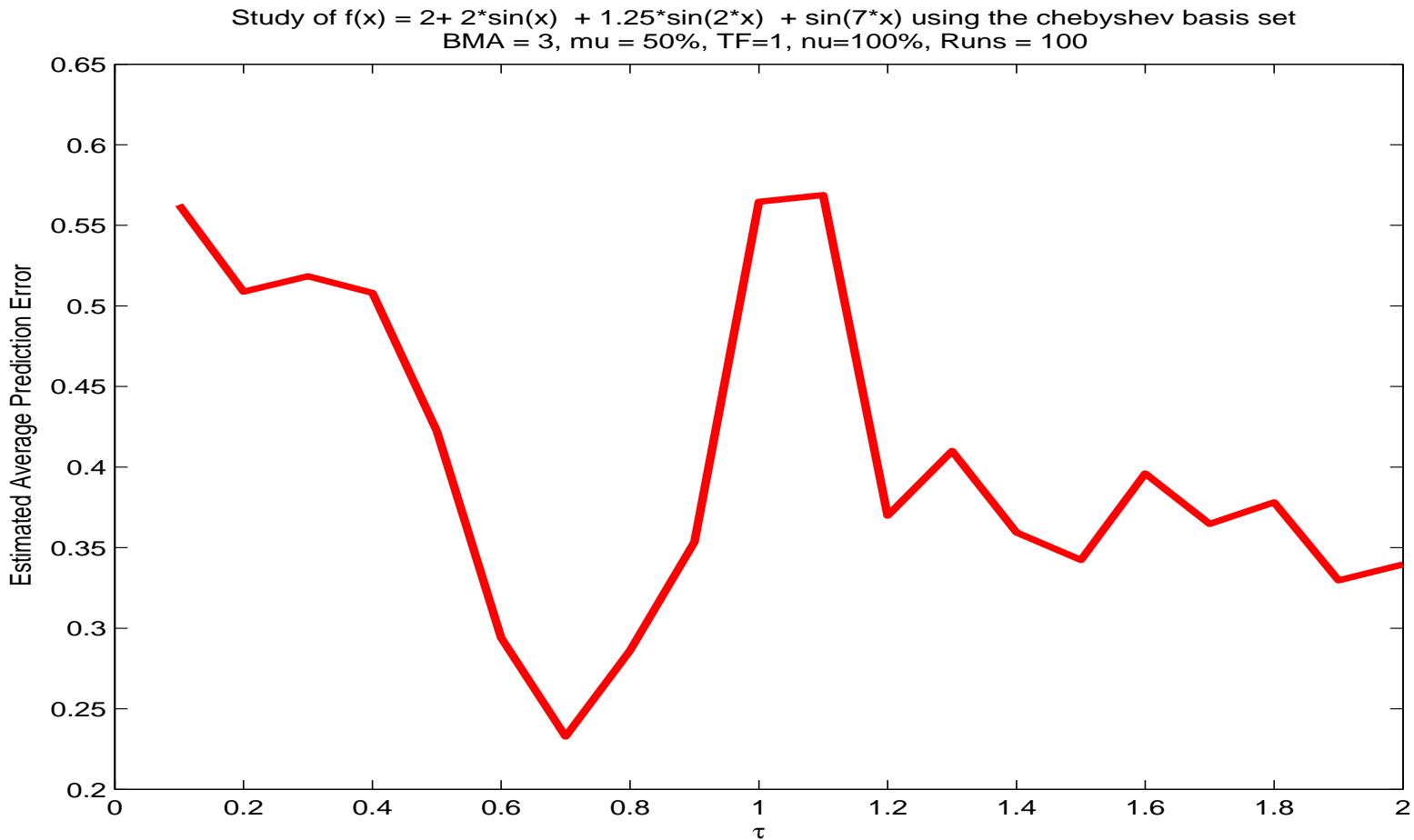
$$p(M|D) \rightarrow 0 \quad \text{as} \quad |\mathcal{M}| \quad \text{gets large}$$

- A remedy: Sequential model list selection

Insights and Conjectures

- *For a given problem and an optimality criterion thereof there must exist an optimum model list.*
- *Such an optimal model list achieves the best bias-variance trade-off for the given problem.*
- *Regularization in model space*

Evidence of optimum model list



The x-axis on the above graph is a model list index. Clearly, we see that there is an optimum model list at 0.7.

Sequential Model List Selection

The building blocks of the method are:

- Selection threshold τ where $\tau \in [0, 2]$
- Working basis set $\mathcal{W}^{(t)} \subseteq \mathcal{E}$
- Term formation scheme (TF)
- Averaging scheme (BMA)
- Proportion of terms to use ($\nu \in [0, 1]$)
- Proportion of models to include ($\mu \in [0, 1]$)
- Distance measure $d(\cdot, \cdot)$ use to search \mathcal{E}

Remember that our **goal is predictive optimality**

Model Averaging Schemes

What models go into the average?

We use an index named BMA to identify the scheme

- BMA = 1
Small size models: 1, 2, 3 terms in the models
- BMA = 2
Medium size models: $p/2$ terms in the models
- BMA = 3
Large size models: $p, p-1, p-2$ terms in the models

Note: For a given scheme, the selection randomly draws $100\mu\%$ of the models available in the induced space.

Term Formation Schemes

Motivation: Terms formed as a combination of atoms from the basis set \mathcal{E} tend to produce sparse function approximations.

- TF = 1

Use $\mathcal{B}^{(t)} = \mathcal{W}^{(t)}$ directly without any partial sums.

- TF = 2

$\mathcal{B}^{(t)} = \{ \text{Partial sums of two elements from } \mathcal{W}^{(t)} \}$

- TF = 3

$\mathcal{B}^{(t)} = \{ \text{Partial sums of three elements from } \mathcal{W}^{(t)} \}$

For a given TF, randomly draw $100\nu\%$ of the terms.

Useful for assessing the efficacy of overcompleteness?

Function Approximation

At time point t

- Get $D^{(t)} = \{(\mathbf{x}_i, y_i), i = 1, \dots, mt\}$
- Construct $\text{BMA}^{(t)}$ using BMA and TF
- Estimate the response for $D^{(t)}$:
 - $\hat{y}_i = \text{BMA}^{(t)}(\mathbf{x}_i)$
- Compute the first order residuals:
 - $r_i = y_i - \hat{y}_i = y_i - \text{BMA}^{(t)}(\mathbf{x}_i)$

Update the Model List

- Search $\mathcal{E} \setminus \mathcal{W}^{(t)}$

for $j = 1$ to $|\mathcal{E} \setminus \mathcal{W}^{(t)}|$

$$\mathbf{r} := (r_1, r_2, \dots, r_{mt})^\top$$

$$\mathbf{e}_j := (\mathbf{e}_j(\mathbf{x}_1), \mathbf{e}_j(\mathbf{x}_2), \dots, \mathbf{e}_j(\mathbf{x}_{mt}))^\top$$

$$\rho_j := d(\mathbf{e}_j, \mathbf{r})$$

if $\rho_j \leq \tau$ then $\mathcal{W}^{(t)} := \mathcal{W}^{(t)} \cup \{\mathbf{e}_j\}$

end

Automation of residual analysis.

What distance to use?

- Norm

$$d(\mathbf{e}_j, \mathbf{r}) := \left\| \frac{\mathbf{e}_j}{\|\mathbf{e}_j\|} - \frac{\mathbf{r}}{\|\mathbf{r}\|} \right\|$$

- Inner Product

$$d(\mathbf{e}_j, \mathbf{r}) := g \left(\left\langle \frac{\mathbf{e}_j}{\|\mathbf{e}_j\|}, \frac{\mathbf{r}}{\|\mathbf{r}\|} \right\rangle \right)$$

- Similarity measures (kernel).

$$d(\mathbf{e}_j, \mathbf{r}) := K(\mathbf{e}_j, \mathbf{r})$$

Some important issues

- Allow only the best candidate
 - Parsimony of model list
 - Not computationally efficient
- Allow all the good guys
 - Allows "not so good" guys
 - More computationally efficient
- Consider stochastic search schemes

Sequential Model List Selection

- For a $\tau \in [0, 2]$
- At time point t
 - Receive m i.i.d observations.
 - Get working set $\mathcal{W}^{(t)} \subseteq \mathcal{E}$.
 - Form term set $\mathcal{B}^{(t)}$ from $\mathcal{W}^{(t)}$
 - Form BMA^(t) using $\mathcal{B}^{(t)}$ and typology.
 - Update $\mathcal{W}^{(t)}$ according to τ

Basis Sets Considered

- *Full Fourier basis*

- $\mathcal{E} = \{\sin(j\omega\mathbf{x}), \cos(j\omega\mathbf{x})\}$

- *Legendre*

- $(j + 1)e_{j+1}(\mathbf{x}) = (2j + 1)\mathbf{x}e_j(\mathbf{x}) - je_{j-1}(\mathbf{x})$

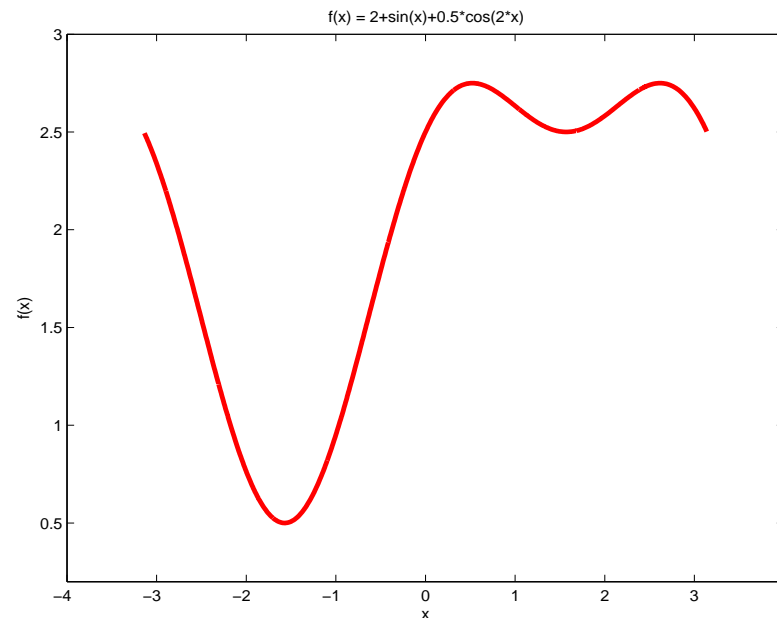
- *Chebyshev*

- $e_j(\mathbf{x}) = \cos(j \arccos(\mathbf{x}))$

- *Fourier sine: $\mathcal{E} = \{\sin(j\mathbf{x})\}$*

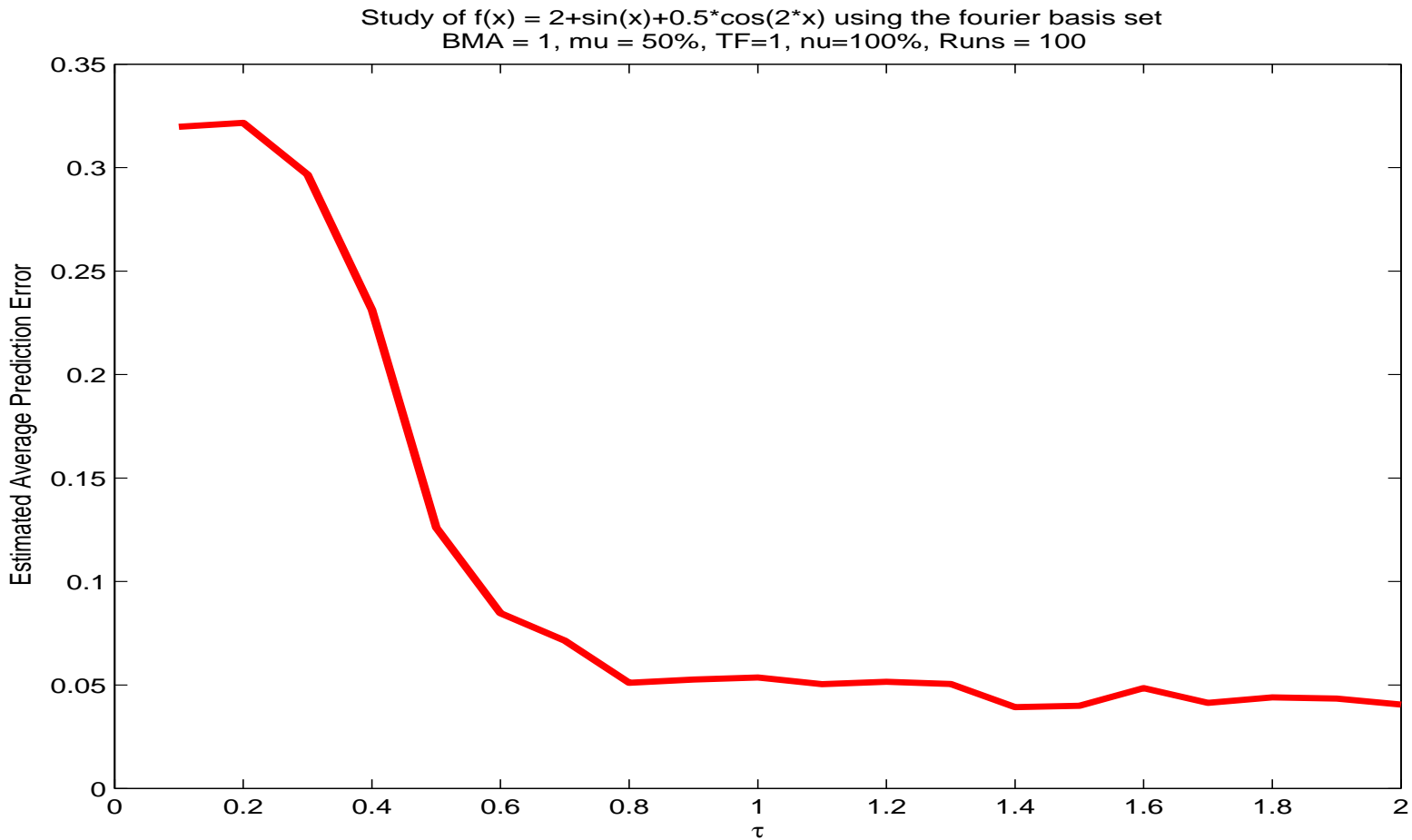
The Deep Valley Function

- $\mathbf{x} \in [-\pi, \pi]$
- $f^*(\mathbf{x}) = 2 + \sin(\mathbf{x}) + 0.5 \cos(2\mathbf{x})$



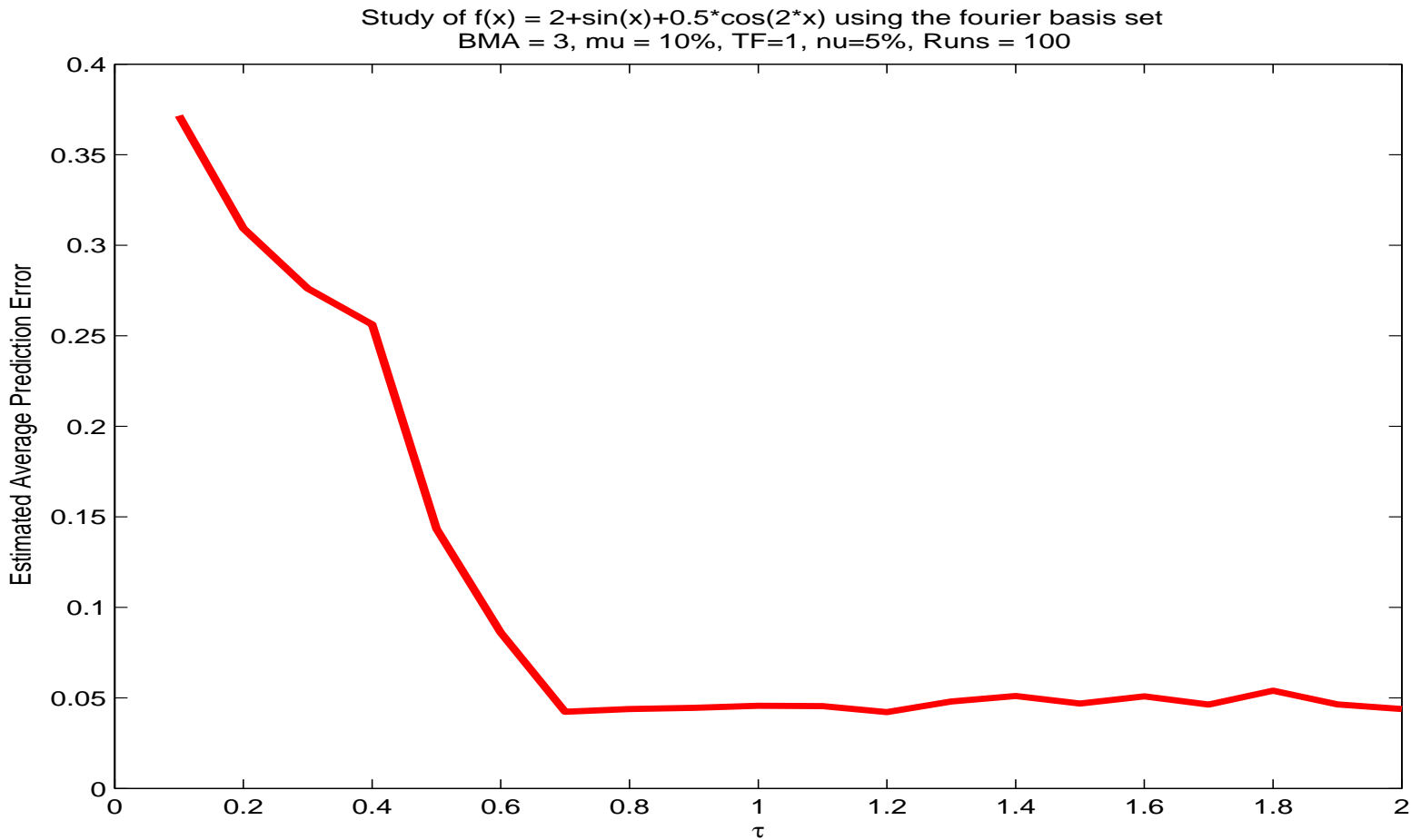
- Study using various \mathcal{E} , τ , TF, BMA, μ and ν .
- **Note: This function lives in the span of Fourier**

Lists with small size models and simple terms when $f^* \in \text{span}\mathcal{E}$



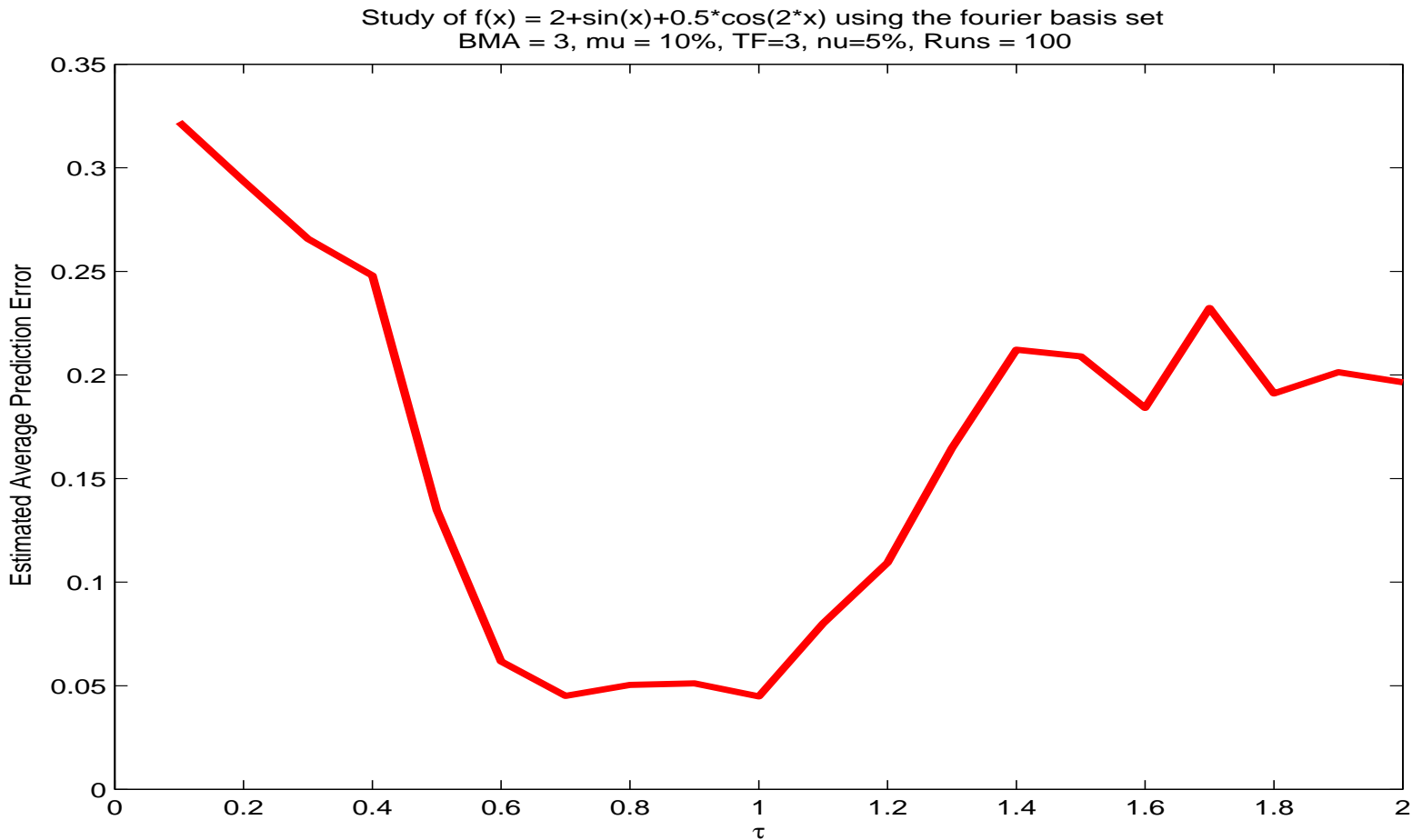
$f^ \in \text{span}\mathcal{E}$: With small size models, the prediction error stabilizes. Not need to add more models beyond the optimum once the optimum model list is found.*

Lists with large size models, and simple terms when $f^* \in \text{span}\mathcal{E}$



$f^ \in \text{span}\mathcal{E}$: With large size models, and simple terms, the prediction error stabilizes. Not need to add more models beyond the optimum once the optimum model list is found.*

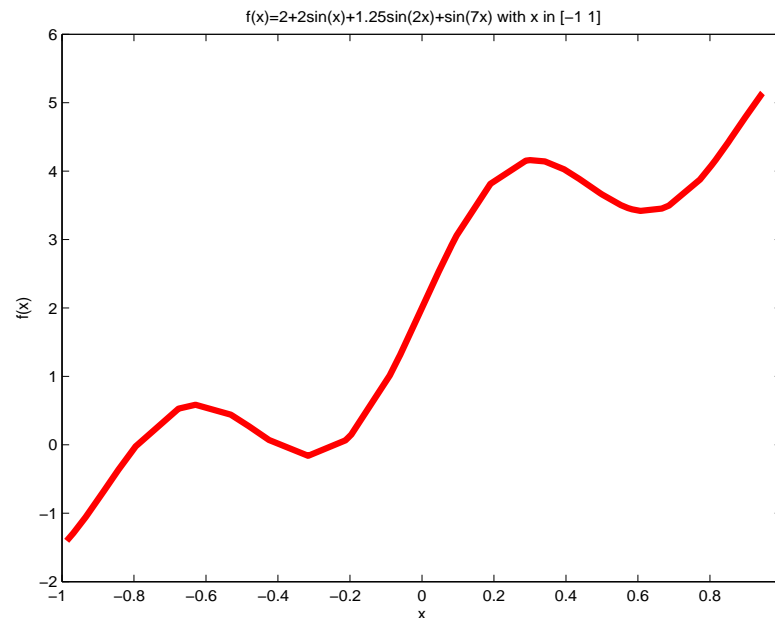
Lists with large size models, and complex terms when $f^* \in \text{span}\mathcal{E}$



$f^ \in \text{span}\mathcal{E}$: With large models, and complex terms, there is a clear optimum. Adding becomes bad beyond the optimum.*

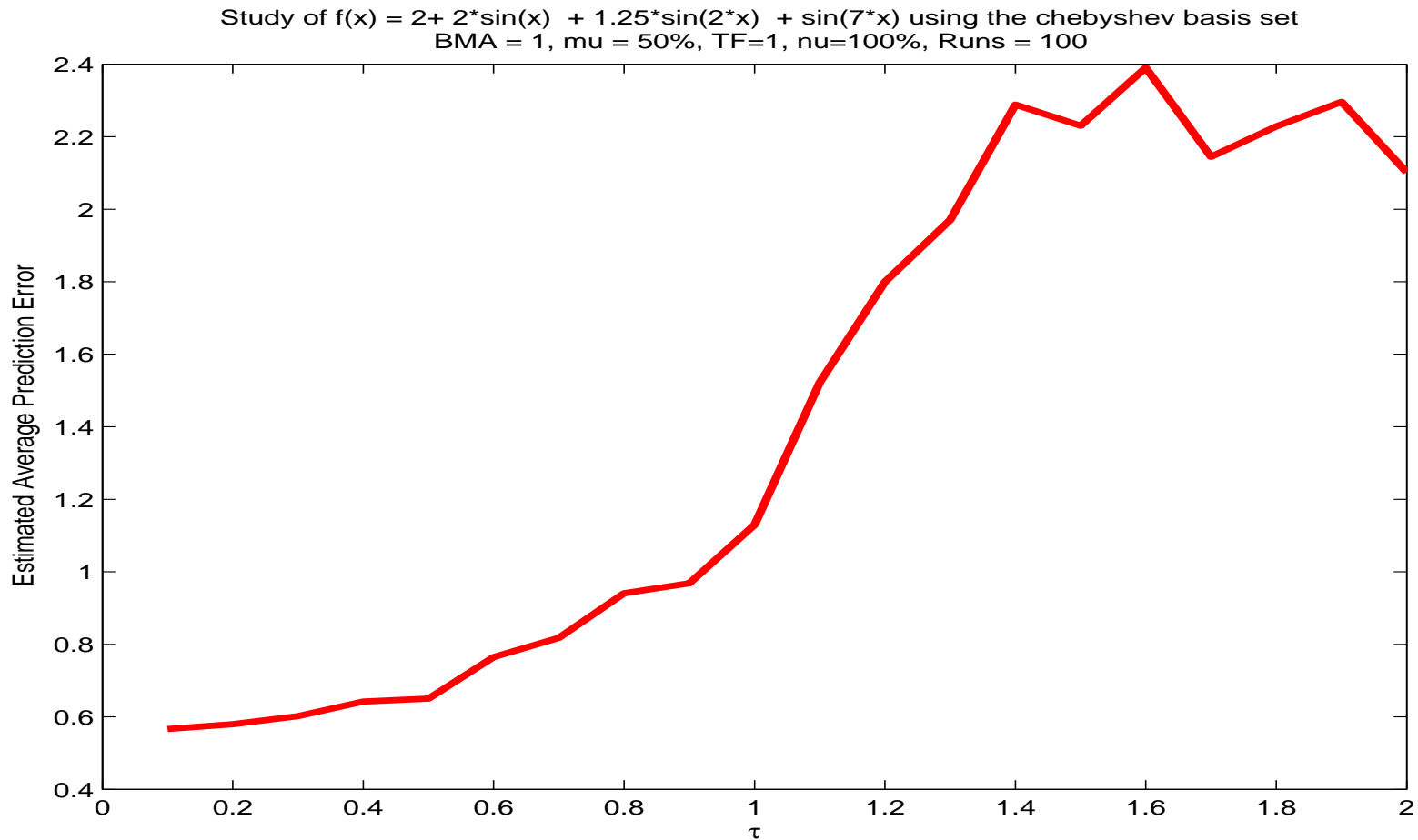
The Nice Hill Function

- $\mathbf{x} \in [-1, 1]$
- $f^*(\mathbf{x}) = 2 + 2 \sin(\mathbf{x}) + 1.25 \sin(2\mathbf{x}) + \sin(7\mathbf{x})$



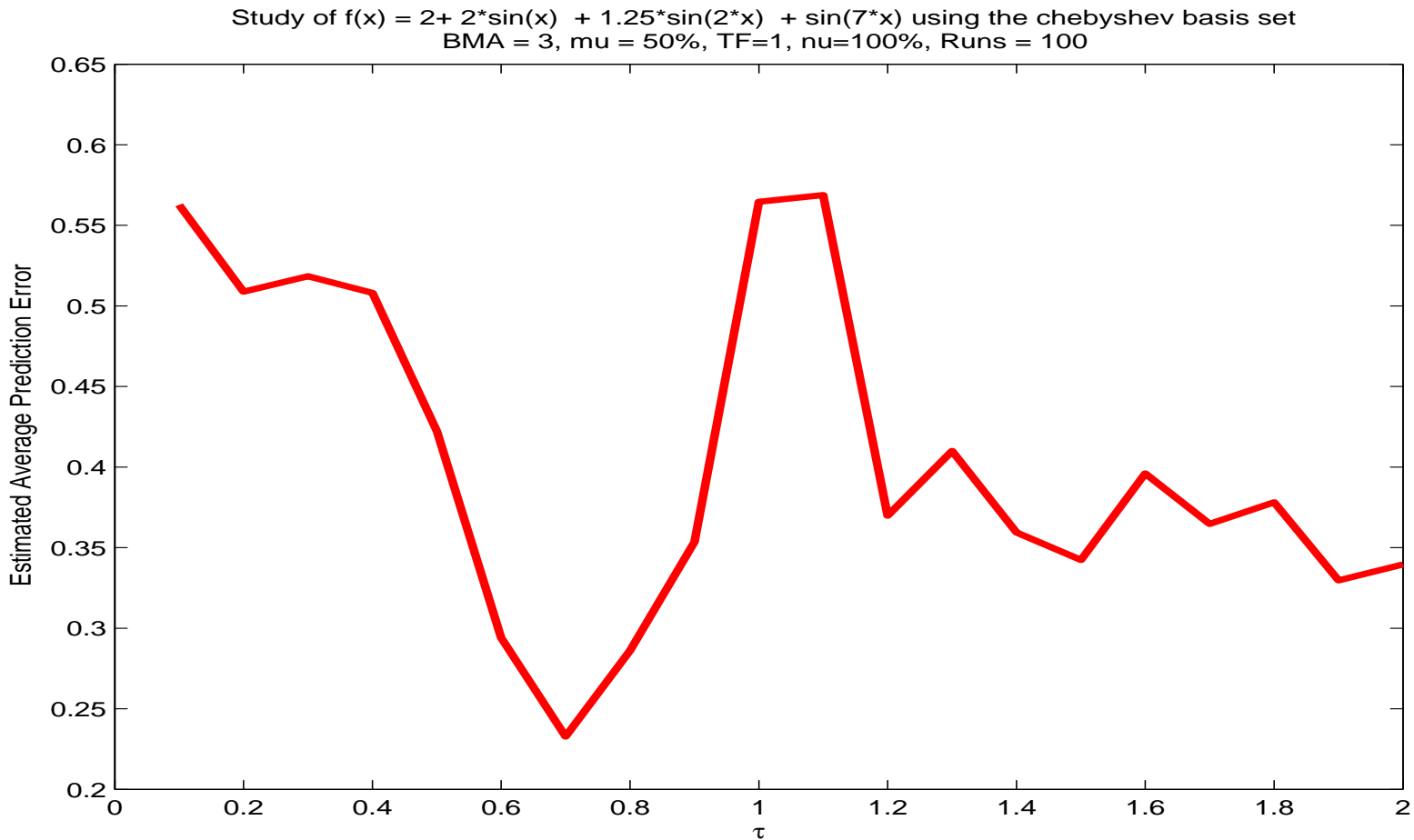
- Study using various \mathcal{E} , τ , TF, BMA, μ and ν .

Lists with small size models, and simple terms when $f^* \notin \text{span}\mathcal{E}$



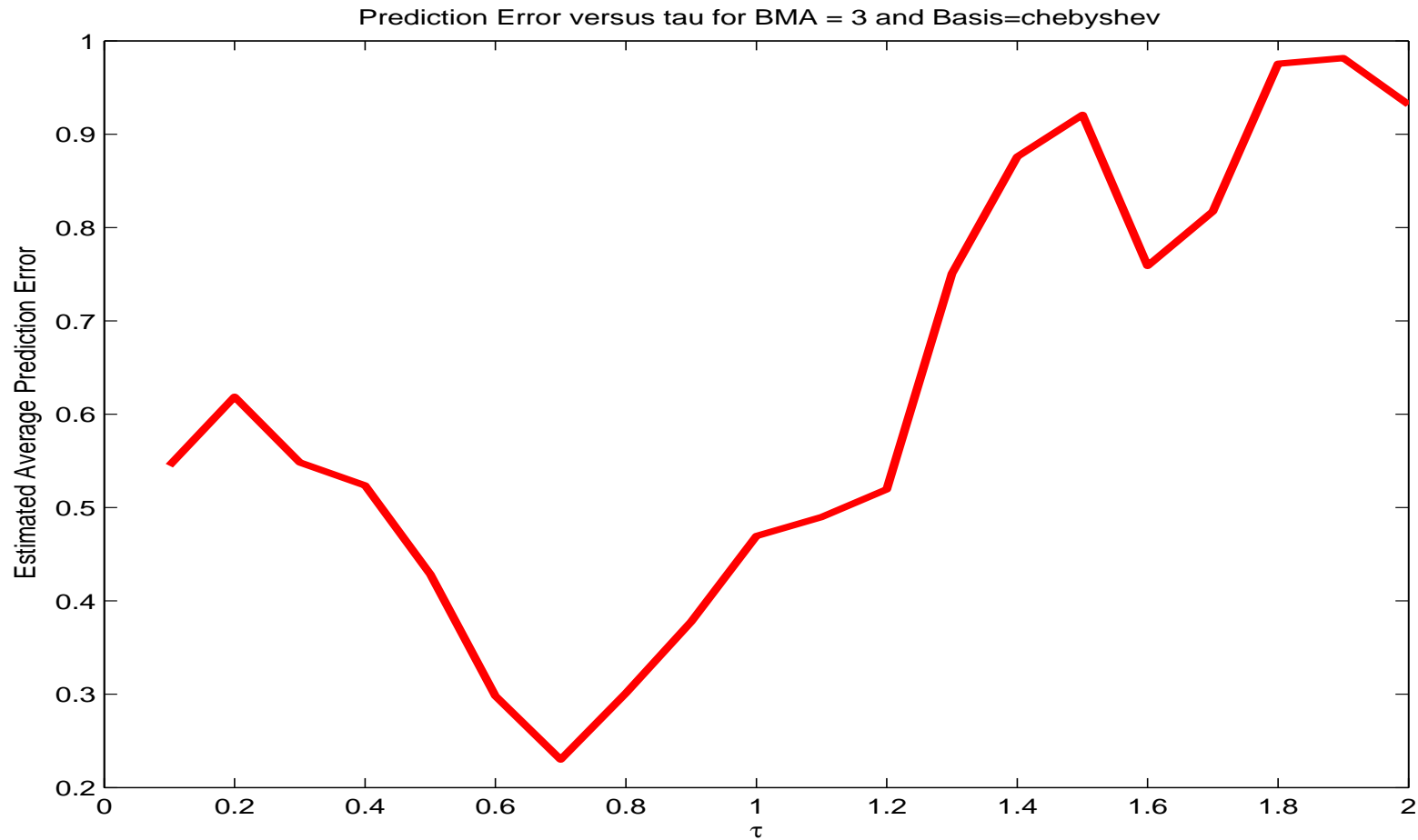
$f^* \notin \text{span}\mathcal{E}$: With small size models, and simple terms, the prediction error increases dramatically with τ . Model lists should be kept small in such a case.

Lists with large size models, and simple terms when $f^* \notin \text{span}\mathcal{E}$



$f^* \notin \text{span}\mathcal{E}$: With large size models, and simple terms, there is a clear optimum model list $\mathcal{M}^*(\tau)$.

Lists with large size models, and complex terms when $f^* \notin \text{span}\mathcal{E}$



$f^* \notin \text{span}\mathcal{E}$: With large size models, and complex terms, there is a clear optimum model list $\mathcal{M}^*(\tau)$.

What to make of all that?

Emerging trends

- $\mathcal{M} = \{\text{small size models}\}$ then $f^* \in \text{span}\mathcal{E}$ there is an optimum beyond which there is neither improvement nor deterioration.
- $\mathcal{M} = \{\text{large size models}\}$ then there seems to be a clear optimal model list.
- $\mathcal{M} = \{\text{small size models}\}$ then if $f^* \notin \text{span}\mathcal{E}$ large model lists are not good

Conclusion and Future

● Take this home

- Model List Variability is indeed one of the main sources of uncertainty.
- Model List Selection is therefore of vital importance.

● Future work

- How does one go about estimating the optimum τ in real settings?
- Test on real data and multivariate predictors.
- Explore more theoretical aspects.