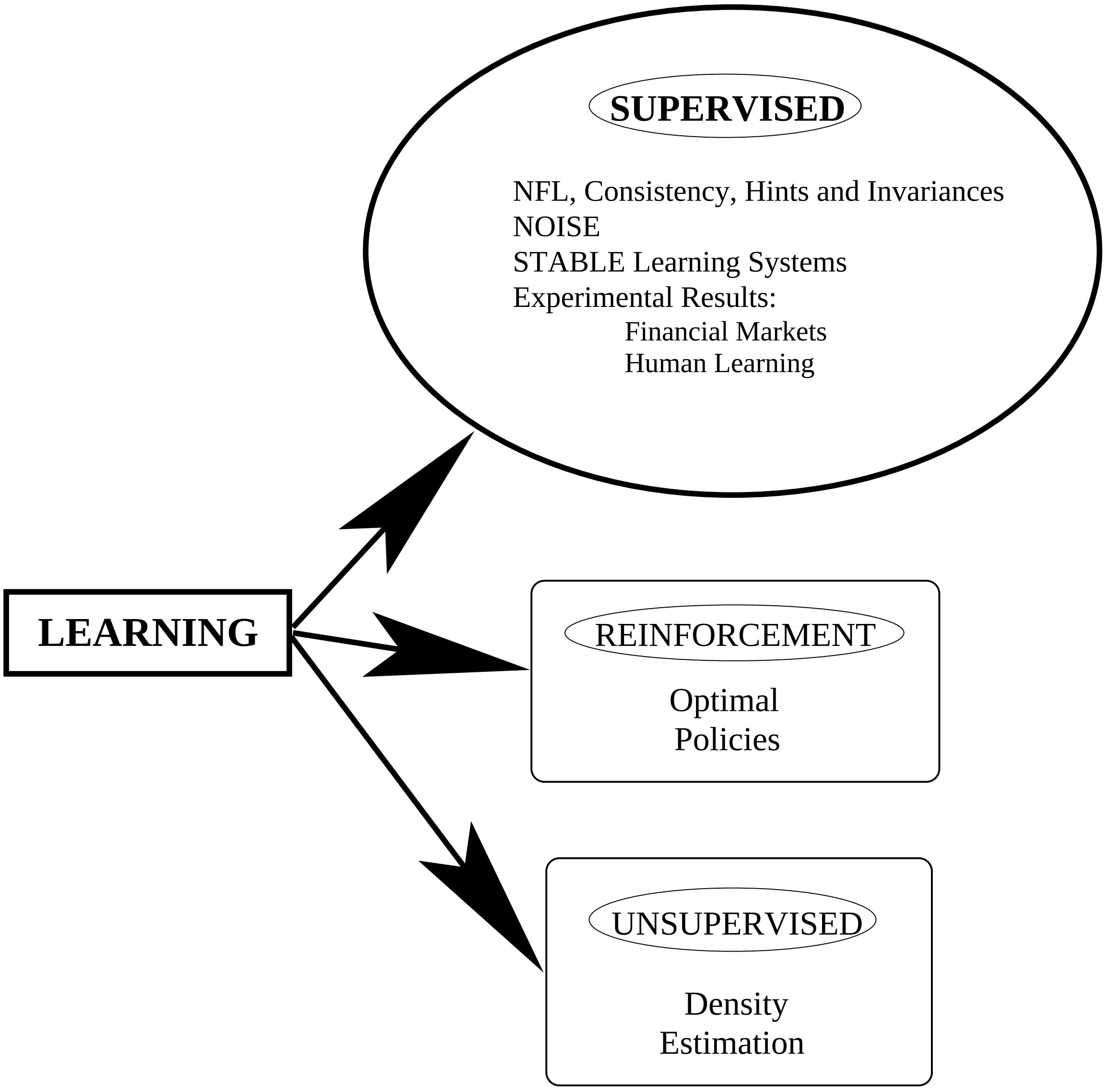
EFFECTIVE CHOICE OF LEARNING SYSTEMS IN THE PRESENCE OF NOISE

Malik Magdon-Ismail
Learning Systems Group, Caltech.
and
Machine and Computational Learning Group, RPI.

Yaser S. Abu-Mostafa
Learning Systems Group, Caltech.

Alexander Nicholson
Learning Systems Group, Caltech.

OVERVIEW


SUPERVISED
NFL, Consistency, Hints and Invariances
NOISE
STABLE Learning Systems
Experimental Results:
Financial Markets
Human Learning
LEARNING
REINFORCEMENT
Optimal
Policies
UNSUPERVISED
Density
Estimation

What is Learning?

Learning is a general concept that describes the process of formulating a *hypothesis* from a finite set of *data*.

Example: Distinguishing between Bach/Beethoven

> **data**: a set of pieces, Bach and Beethoven.
>
> **learning**: develop some method for differentiating between Bach and Beethoven.

In general, a **data set** is composed of a set of **instances** that represent the learning task. Learning entails choosing a single solution from an available set of "task solvers". The ultimate performance is the test of this solution on unseen data.

Generality of Learning

- *Probability Density Estimation*. The instances are data drawn from the true distribution.

- *Pattern Recognition*. The instances are a pair {feature vector, class label}.

- *Regression*. The instances correspond to $\{\mathbf{x}, y\}$ pairs and the underlying process which is to be learned is the mapping $y = f(\mathbf{x})$.

- *Time Series Prediction*. The data usually consists of a single time series. The task is to predict the next point.

NOISE

1. You are sitting there trying to learn to recognize the difference between Bach and Beethoven and your little brother is running around screaming at the top of his lungs.

2. We would like to learn to predict the expected college GPA from SAT scores. The observed GPA's are the expected GPA's plus some random component.

Noise represents the **uncertainty inherent in the learning problem**. We expect our **performance** to be affected by the noise.

NOISE affects Learning Performance

- Corrupts the data. The “learned” hypothesis will be worse.
- Sets a lower bound for the eventual performance.
- In choosing the **learning system**, one should not ignore the possibility of noise in the data.

Choice of Learning System is Important

- **No Free Lunch (NFL)** (Wolpert, 1996): Cannot assume that the learning problem is "random" as then there is nothing you can do.
  - **Hints, Invariances and Prior Information** (Abu-Mostafa, 1990) should be incorporated into the learning system to improve performance.

- **Consistency of the Learning Process** (Vapnik and Chervonenkis, 1971): Certain learning systems allow one to infer performance from learning data statistics.
  - Structural Risk Minimization for choice of learning systems
  - Support Vector Method

Should the fact that noise is present affect our choice of learning system?

Impact of Noise on Learning

- Framework for Learning.
- Systems that are stable in the presence of noise.
- Effect of Noise on Learning Performance.
- Some Experiments.

FORMAL SETUP

The **training data** consists of N data points

$$y_i = f(\mathbf{x}_i) + n_i$$

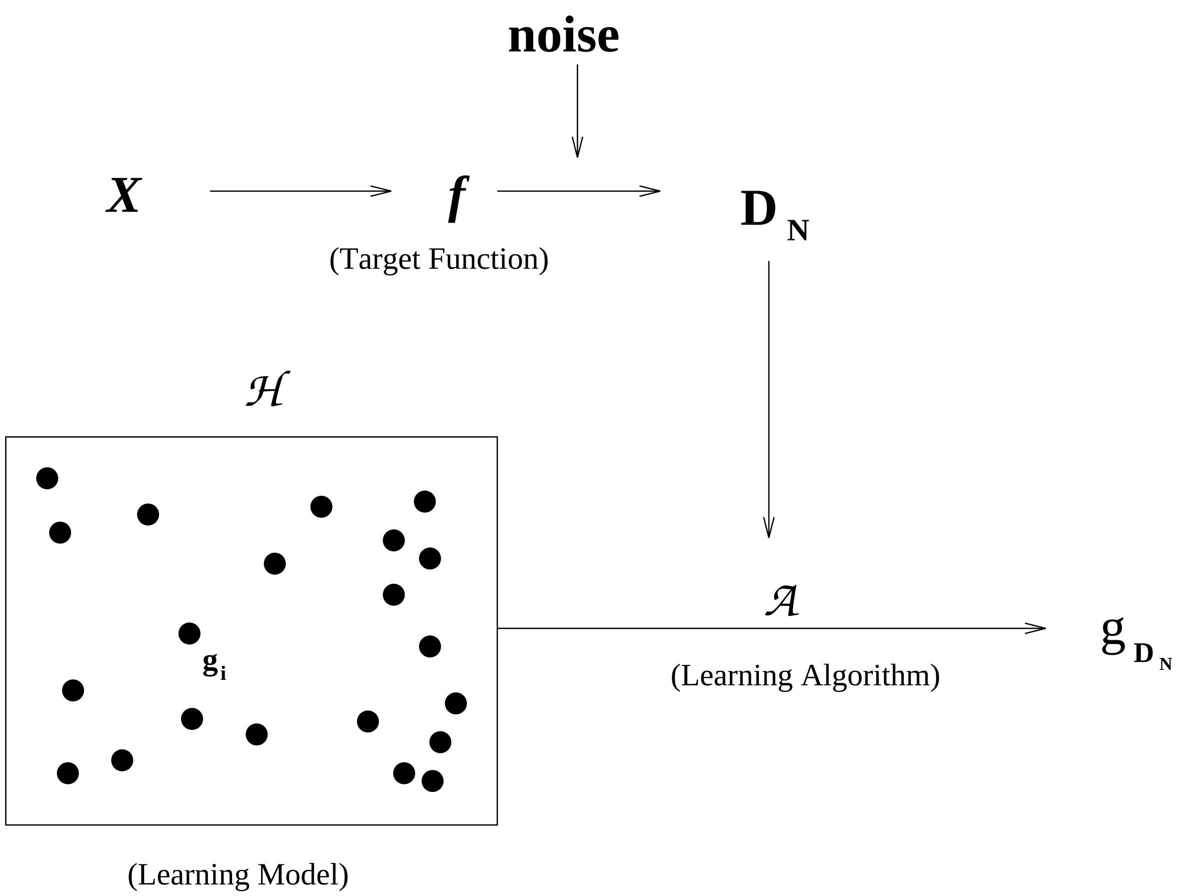
each n_i is an independent noise realization with variance σ_i^2. The average noise variance is $\overline{\sigma^2}$.

The **learning model** $\mathcal{H}$ is a collection of functions that reflects how we choose to model the **target function** f.

The **learning algorithm** is a mapping

$$\mathcal{A} : \mathcal{D}_N \longrightarrow g_{\mathcal{D}_N} \in \mathcal{H}$$

LEARNING SETUP



Performance Measure

For on an unseen data point, $\mathbf{x}$,

$$e(\mathbf{x}) = (g_{\mathcal{D}_N}(\mathbf{x}) - y(\mathbf{x}))^2$$

Which $\mathbf{x}$? Take the expectation. Define the **test error** for $g_{\mathcal{D}_N}$:

$$e(g_{\mathcal{D}_N}) = E\left[g_{\mathcal{D}_N}(\mathbf{x}) - y(\mathbf{x}))^2\right]_{\mathbf{x}}$$

Define the **expected test error**:

$$\mathcal{E}_N(\sigma) = E\left[e(g_{\mathcal{D}_N})\right]_{\mathcal{D}_N}$$

$\mathcal{E}_N(\sigma)$ is a quantitative measure of the performance of a learning system. It depends on the number of data points and the noise level.

EXAMPLE: Zero Dimensional Regression

$$\mathcal{D}_N = \{y_i\}_{i=1}^N$$

$$y_i = m + \sigma\mathcal{N}(0,1), \qquad i = 1 \ldots N$$

I give you N randomly generated numbers and ask you to "learn" the mean.

$$\mathcal{H} : \mathbb{R}$$

An algorithm $\mathcal{A}$ would be method for picking an estimator $\widehat{m} \in \mathbb{R}$.

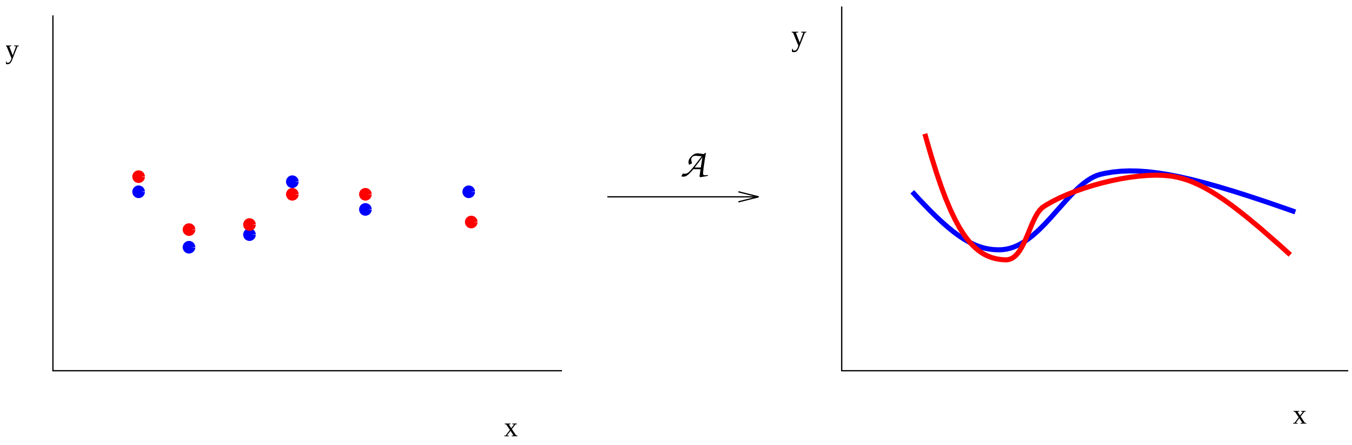
$\mathcal{A}_1$: Choose the $\widehat{m}$ to minimize the sum of squared deviations: $\widehat{m} =$ sample mean. [GOOD]

$\mathcal{A}_2$: Pick $\widehat{m}$ randomly in $\mathbb{R}$. [BAD]

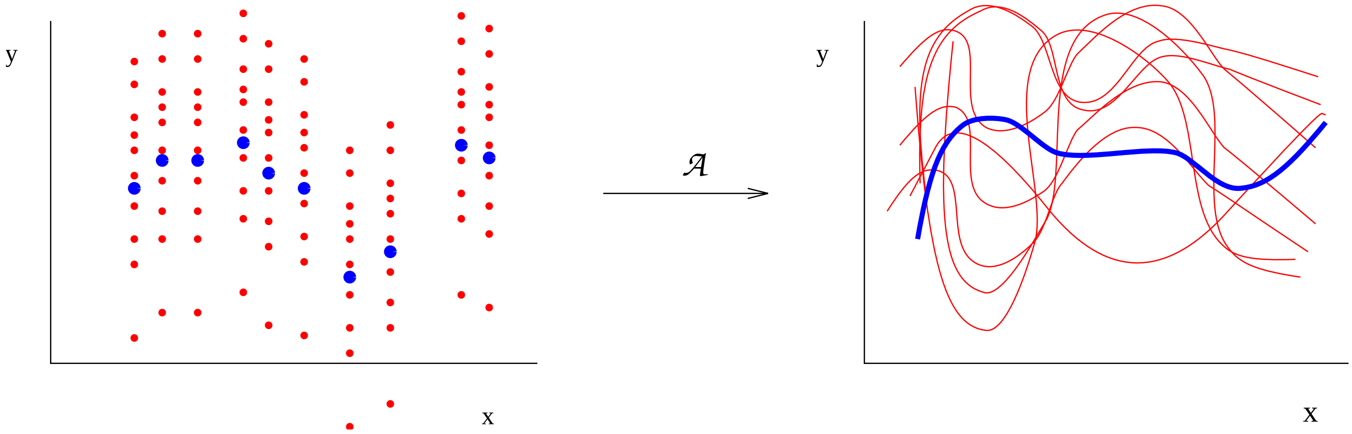
We need some formal definition of a "good" algorithm.

We consider learning systems that are *stable*:

- “Continuous”



- “Mean Preserving”



Stable Learning Systems are "Good"

1. Intuitively plausible.

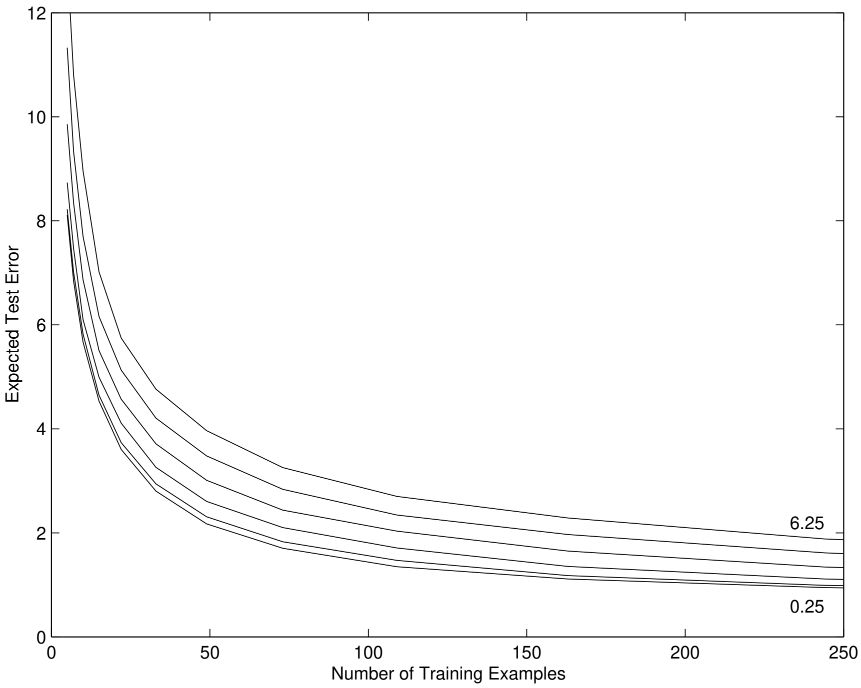
2. Properties of stability have nothing to do with the target function.

3. Can be verified for any given learning system. Linear models, Neural Networks, and RBF's with gradient descent algorithms are stable to within experimental accuracy.

4. Stable learning system performance changes in a continuous way with noise.

5. Even with noise added to every example, the impact of noise on learning diminishes as the data set size increases.

IMPACT OF NOISE ON PERFORMANCE

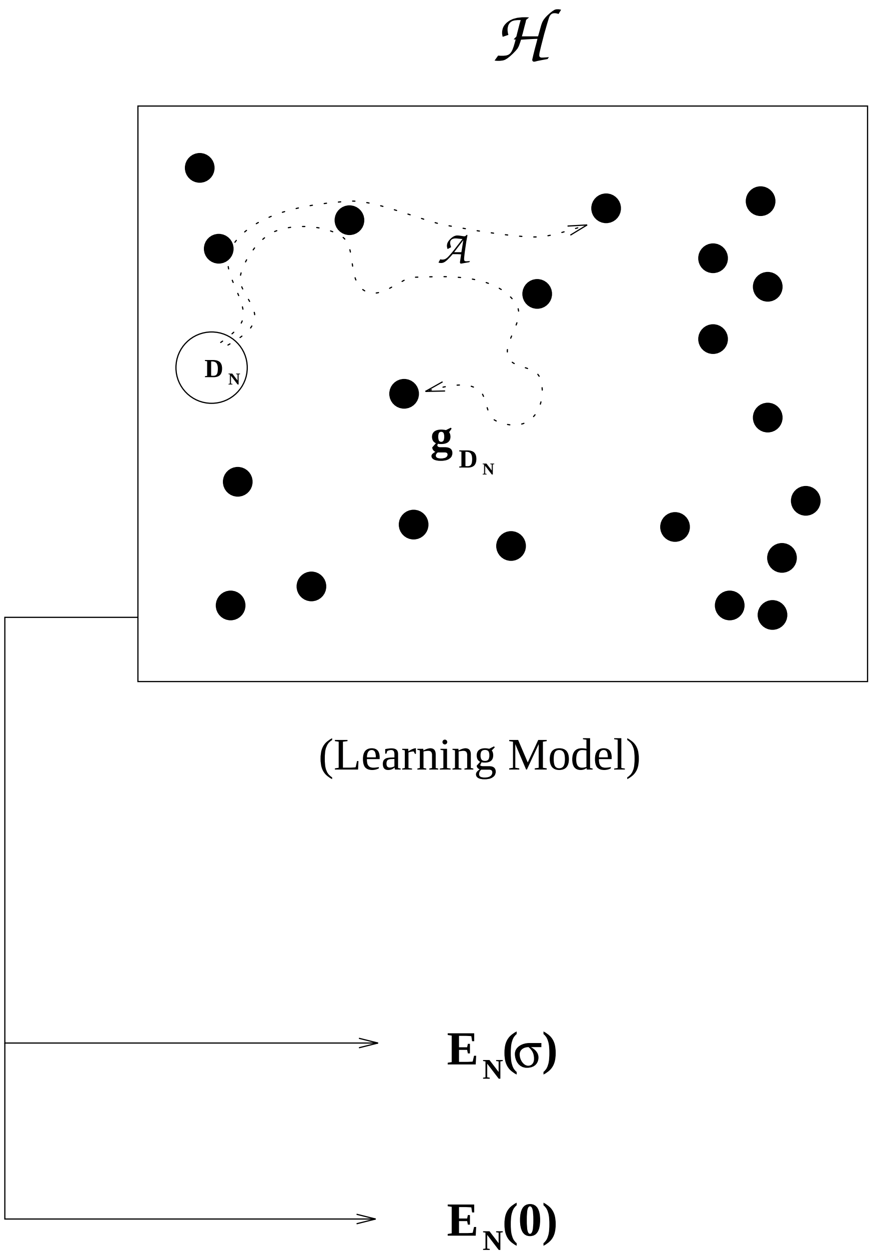


Qualitative observations:

more noise $\Rightarrow$ worse performance

more data $\Rightarrow$ effect of noise decreases
$\Rightarrow$ better performance

Can we quantify this intuition (in the general non linear case)?



How does $E_N(\sigma)$ compare with $E_N(0)$?

Given a learning system, there is some best hypothesis achieving test error E_0.

How does $E_N(0) \to E_0$?

Performance Guarantees

Theorem 1

$$\mathcal{E}_N(\sigma) \leq \mathcal{E}_N(0) + \frac{\overline{\sigma^2} C_1}{N} + o\left(\frac{1}{N}\right) \qquad (1)$$

C_1 is a constant that depends on the learning problem and $\overline{\sigma^2}$ is the average noise variance.

[quantifies the intuition: $noise \uparrow \Rightarrow performance \downarrow$]

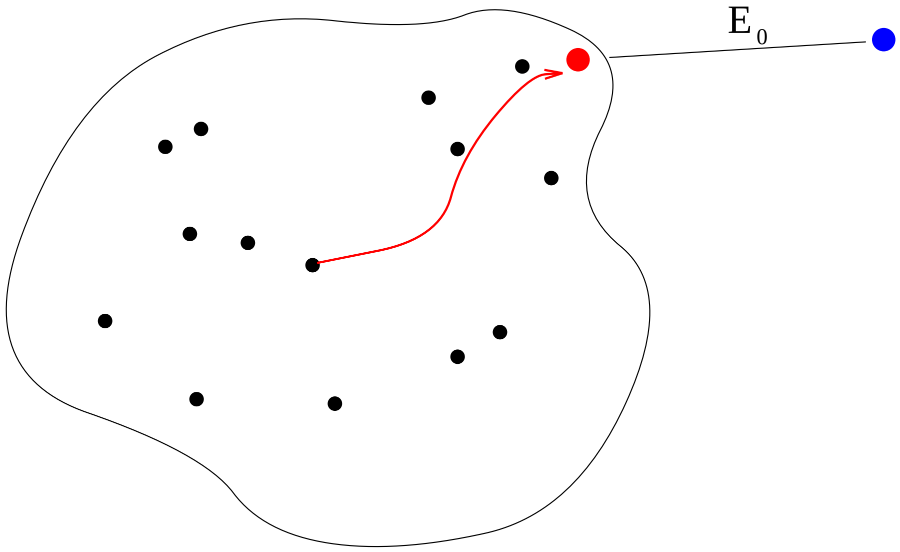
Theorem 2

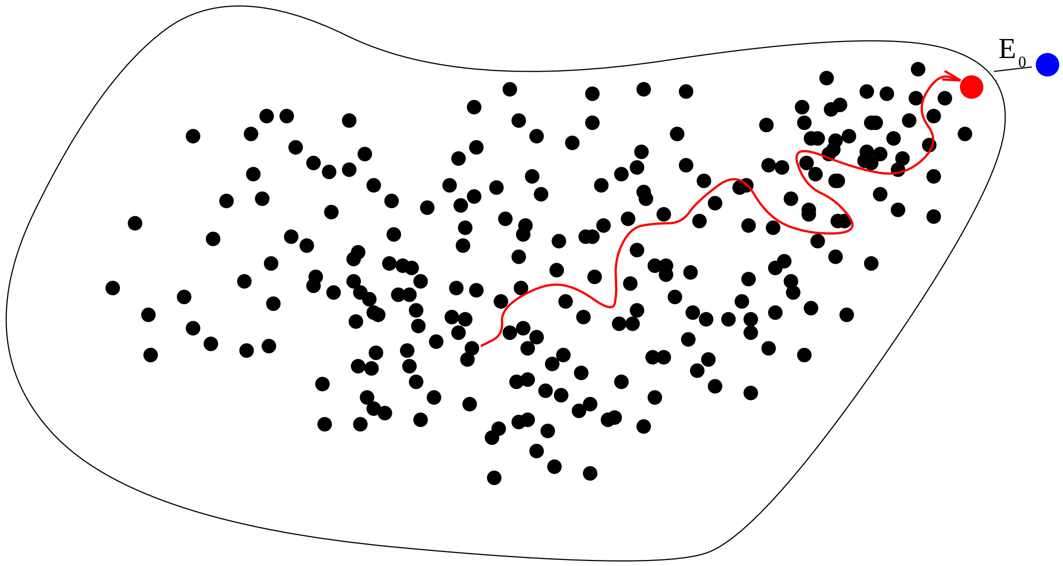
$$\mathcal{E}_N(0) \leq E_0 + \frac{C_2}{N} + o\left(\frac{1}{N}\right) \qquad (2)$$

where $E_0 = \lim_{N \to \infty} \{\mathcal{E}_N(0)\}$ and C_2 is a constant that depends on the learning problem.

[quantifies the performance w.r.t. N.]

Performance / Convergence Rate Trade Off





Comparing to NFL, Hints and VC theory

Stable Learning Systems : $\mathcal{E}_N(\sigma) - \mathcal{E}_N(0)$ is well behaved.

NFL: $\mathcal{E}_N(\sigma)$ interated over a uniform prior over target functions is a constant, independent of choice of learning system.

Hints etc: The prior over target functions is not uniform, and we can potentially both decrease E_0 and the $\mathcal{C}$'s using Hints etc.

VC: For suitable (low VC dim) learning systems, the training performance and test performance are comparable but we dont really know how low either are or how they depend on the noise content in the data.

Experimental Investigation

The theorems imply that

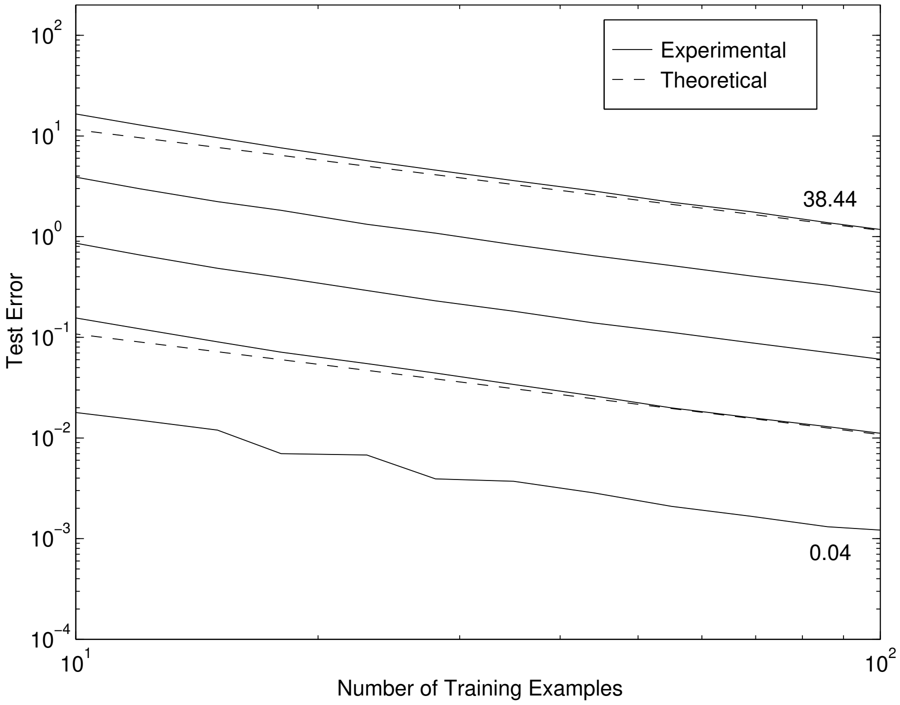
- Error increase is proportional to the noise and inversely proportional to the data set size.

$$\log(\mathcal{E}_N(\sigma) - \mathcal{E}_N(0)) \leq \log \overline{\sigma^2} - \log N$$

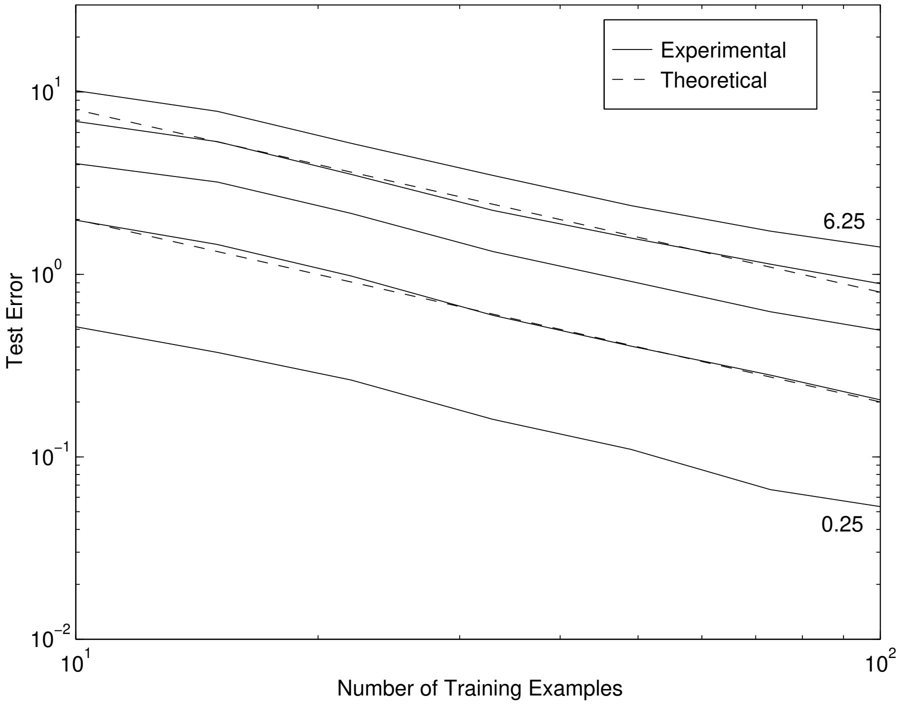
- There is a constant E_0 such that.

$$\mathcal{E}_N(0) - E_0 \leq \frac{C_2}{N}$$

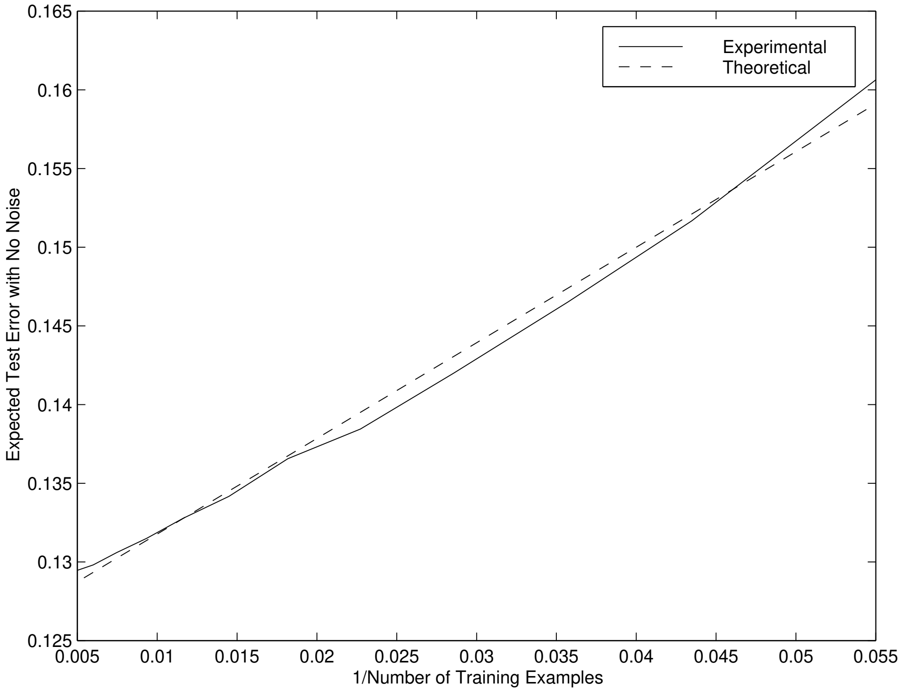
Linear Learning Model



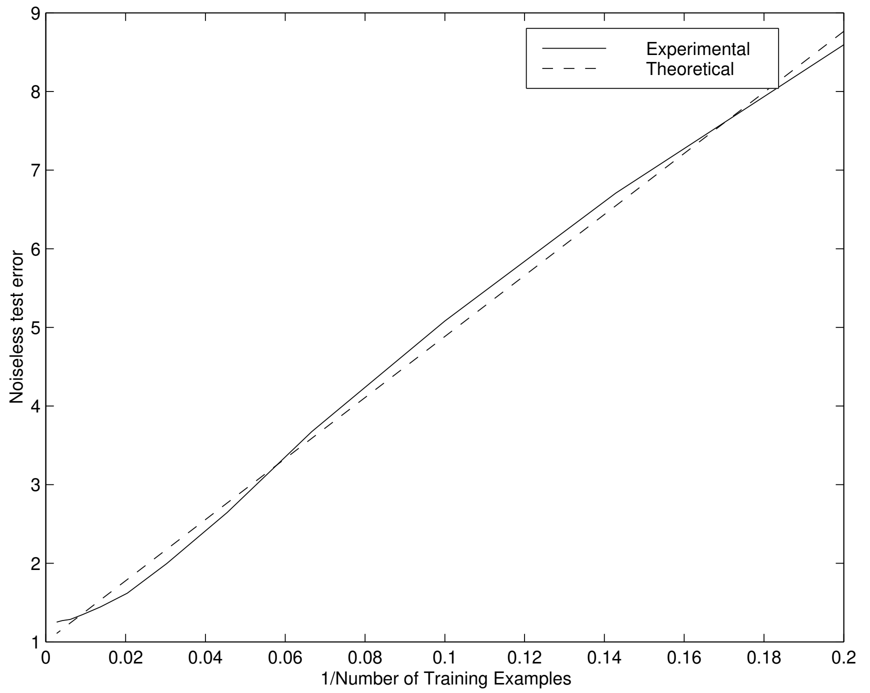
Non-Linear Learning Model



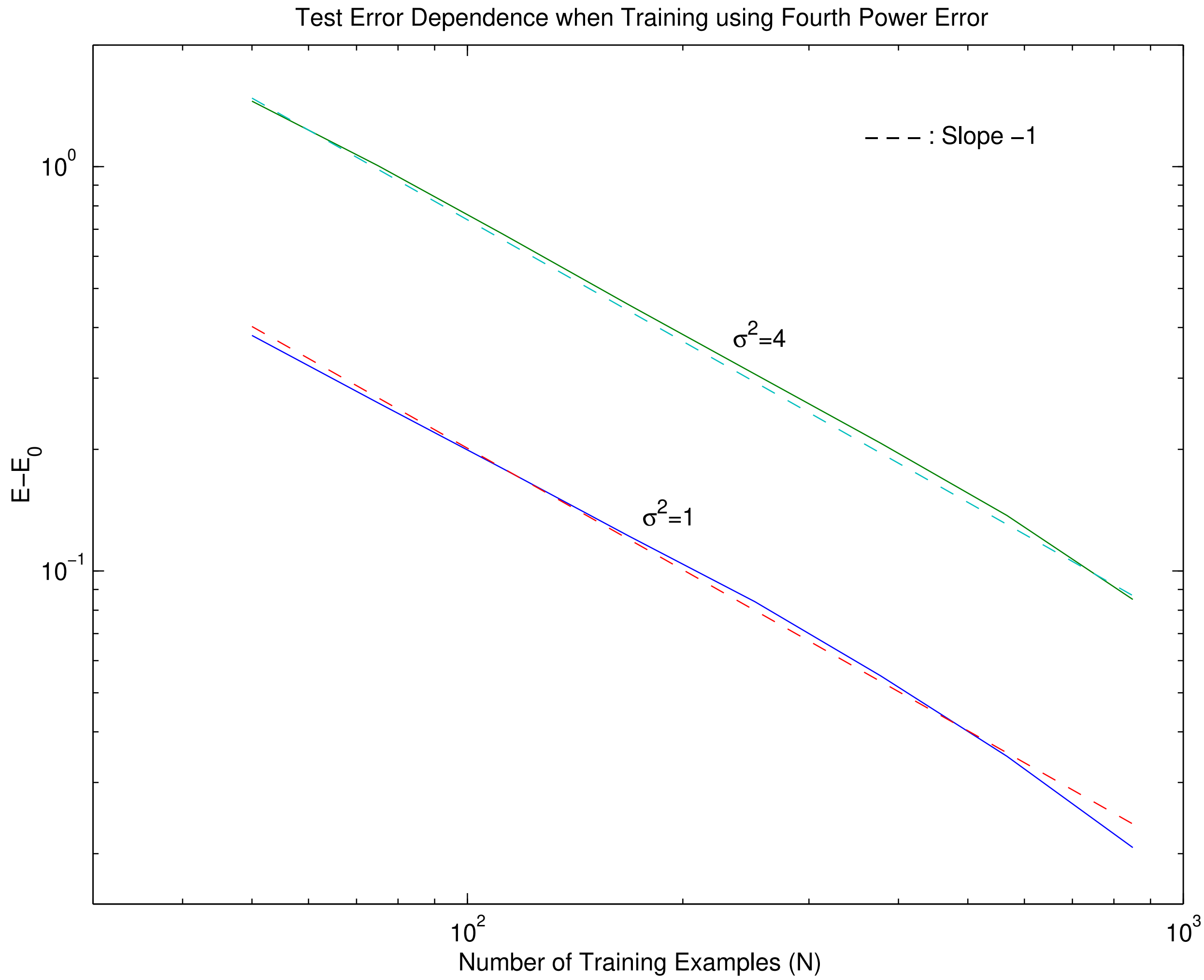
Linear Learning Model



Non-Linear Learning Model



Minimizing the Fourth power of the Deviations (Non-Linear Model)

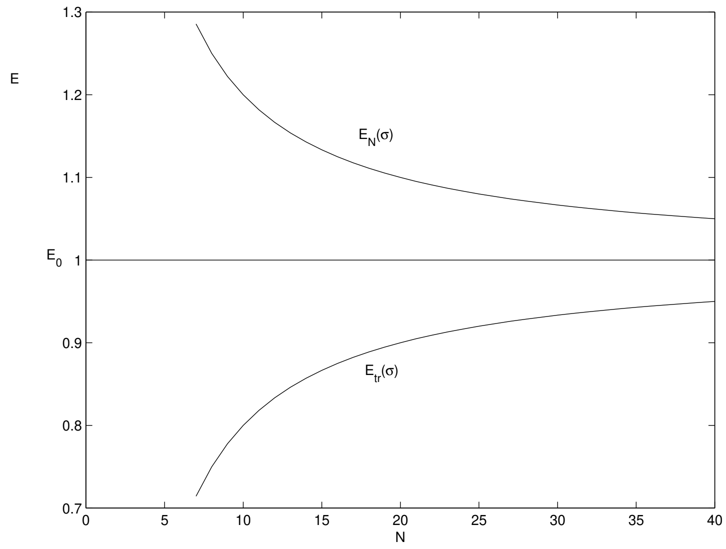


ESTIMATING MODEL LIMITATION

- Experiments show $\mathcal{E}_N(\sigma)$ to quickly assume $1/N$ behavior,

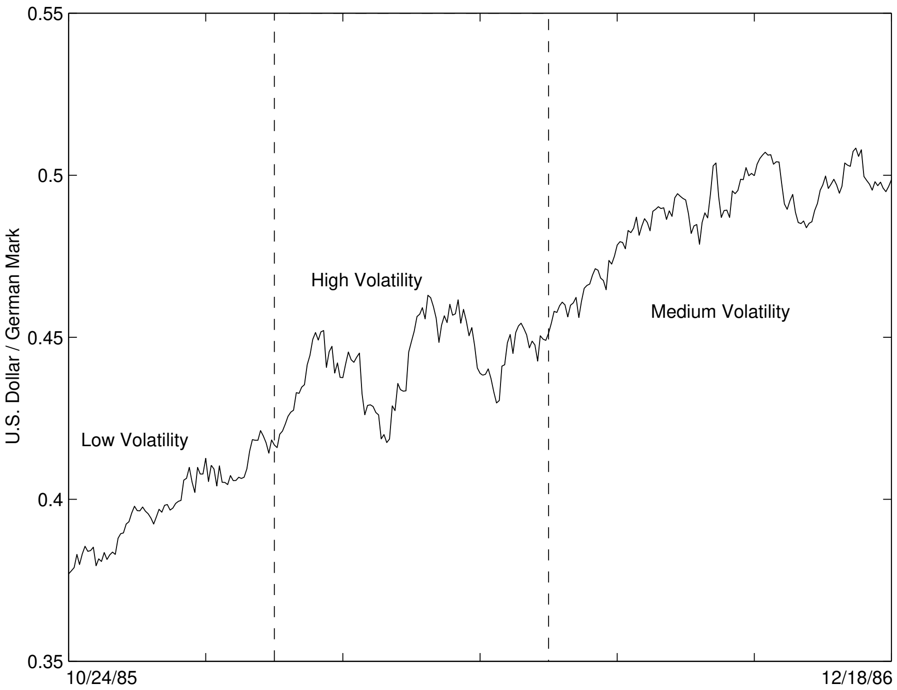
$$E_0 \leq \; \mathcal{E}_N(\sigma) \approx \; E_0 + \frac{\mathcal{C}_1\overline{\sigma^2} + \mathcal{C}_2}{N}$$

(in the case of linear learning models we can replace $\mathcal{C}_1$ by $d+1$).



- Given this functional dependence, can we somehow extract E_0, the model limitation.

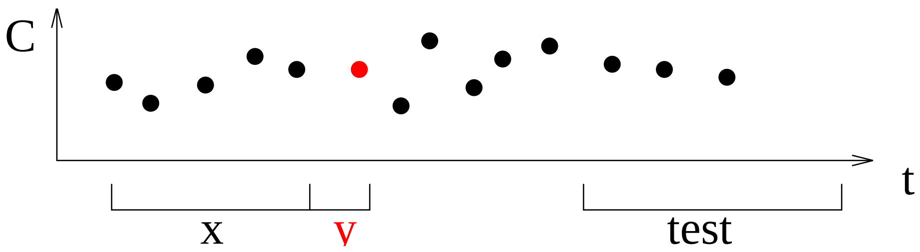
SIMULATION RESULTS IN FINANCIAL MARKETS



We apply the results to the problem of predicting foreign exchange rates with a linear model.

We use the daily close exchange rates 1984-1995 for

Swiss Franc	German Mark
British Pound	Japanese Yen



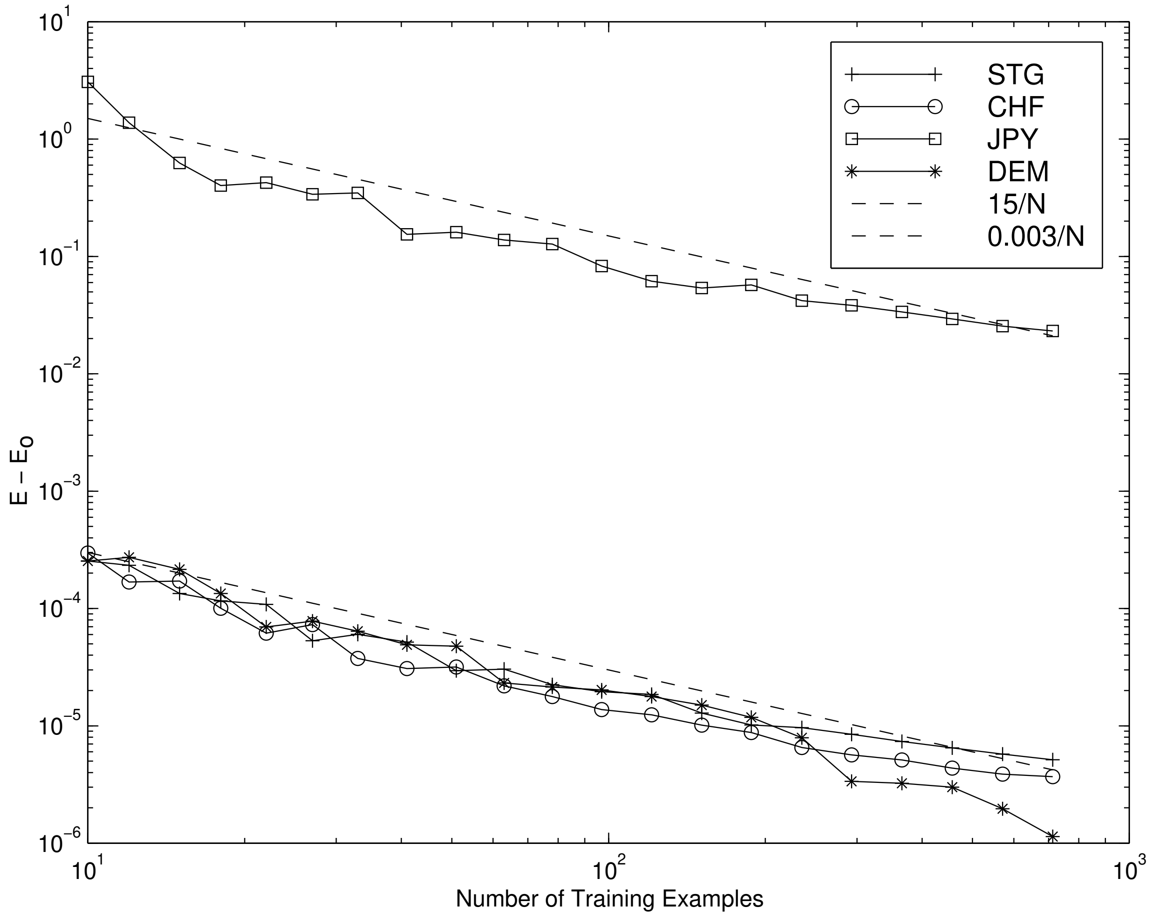
- Create a data set

$$\{\mathbf{x}_k = (S_{k-4}, \ldots, S_k), y_k = S_{k+1}\}$$

leaving out a test set.

- Sample N_1 points repeatedly from this set.

- Vary N_1 and assuming the bounds to be tight, one can fit the observed behavior to the expected behavior.



STG
CHF
JPY
DEM
15/N
0.003/N
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
$E - E_o$
10^{1}
10^{2}
10^{3}
Number of Training Examples

Currency	$E_0 + \overline{\sigma^2}$ Estimate (model lim.)	No Change Predictor (test error)
DEM	0.000499	0.000502
CHF	0.000158	0.000160
STG	0.000134	0.000136
JPY	1.082	1.083

(a)

Currency	$E_0 + \overline{\sigma^2}$ Est. Estimate (model lim.)	No Change Predictor (test error)
DEM	0.000156	0.000152
CHF	0.000148	0.000151
STG	0.000153	0.000157
JPY	0.851	0.867

(b)

In (a) we use the training data alone to estimate E_0*. In (b) we use a separate test set to estimate* E_0*.*

NOISE AND HUMAN LEARNING*

We can perform a similar experiment with human subjects:

- Four groups of human subjects are given data sets of input/output values.

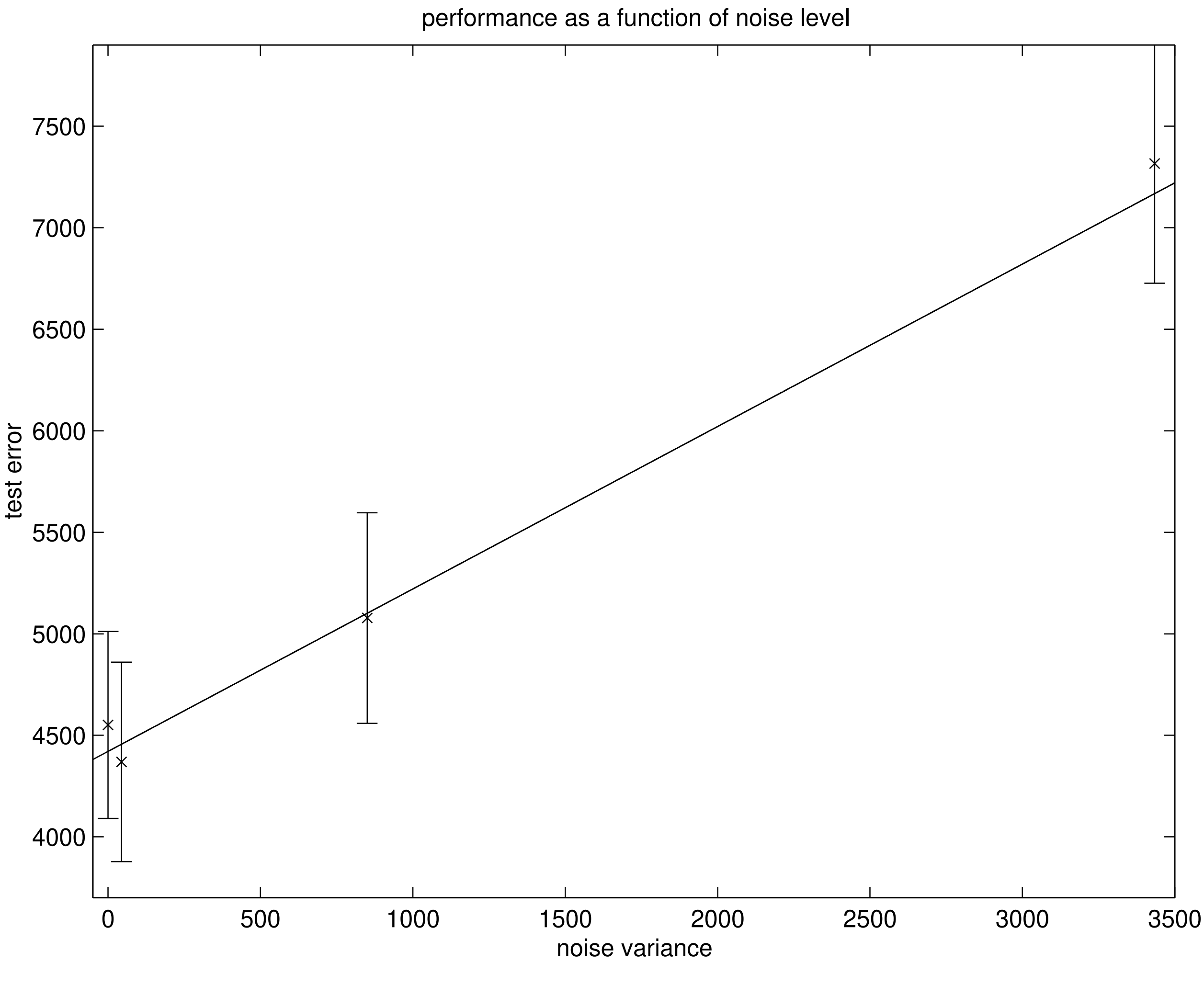
- To the outputs are added noise with differing noise variance for the different groups.

- The goal was to learn the input-output relationship.

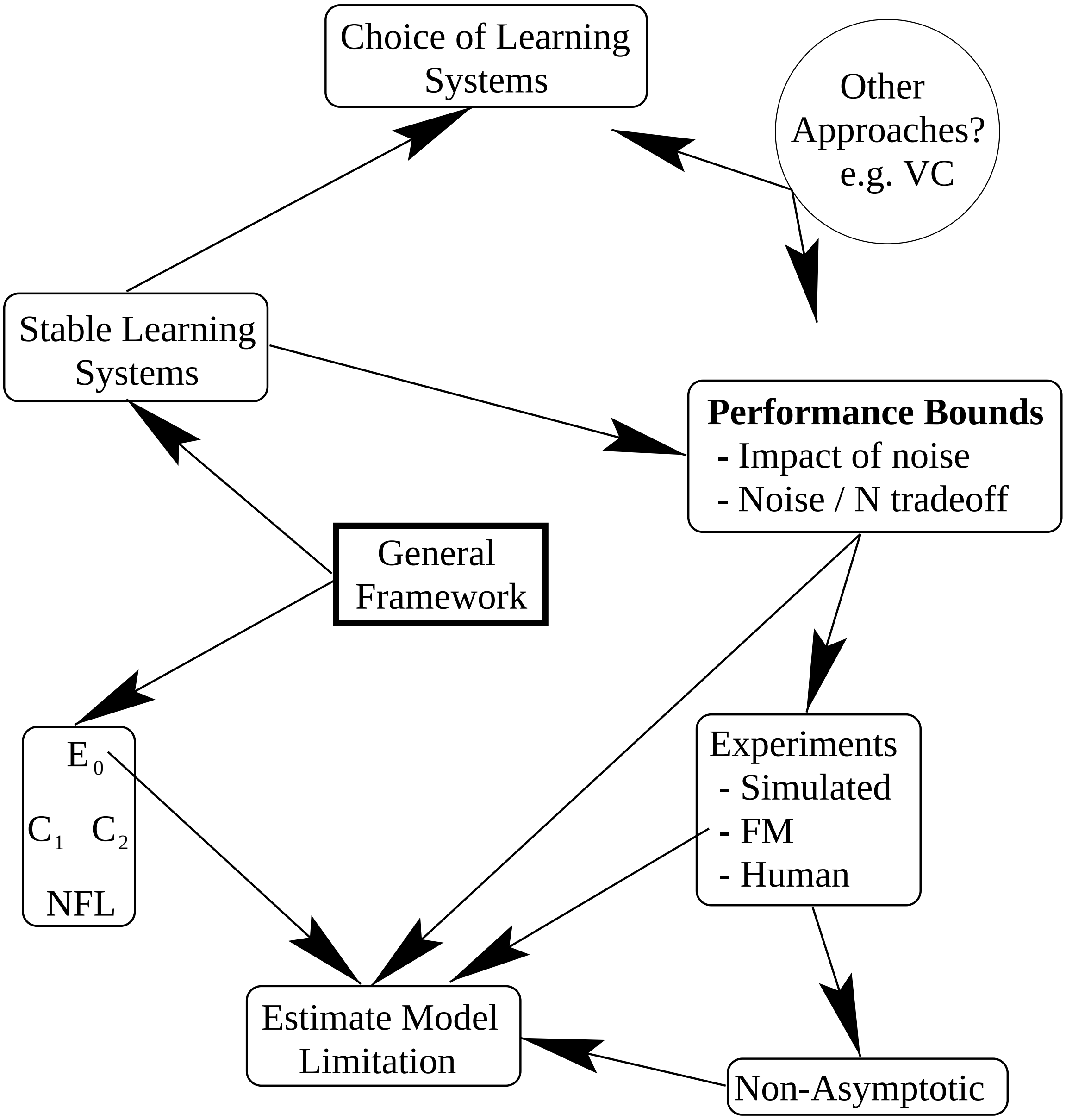
- Performance (MSE) for each subject was tested on an unseen data set.

- We expect the performance error to increase in proportion to $\overline{\sigma^2}$ (Theorem 1)

*These experiments were done in collaboration with Dr. Craig Mckenzie of the UCSD Psychology Dept.


performance as a function of noise level
test error
noise variance
7500
7000
6500
6000
5500
5000
4500
4000
0
500
1000
1500
2000
2500
3000
3500

SUMMARY



Choice of Learning Systems
Other Approaches? e.g. VC
Stable Learning Systems
Performance Bounds
- Impact of noise
- Noise / N tradeoff
General Framework
E_0
C_1 C_2
NFL
Experiments
- Simulated
- FM
- Human
Estimate Model Limitation
Non-Asymptotic

Interesting Questions for Further Investigation

1. Noise in Reinforcement Learning environments. One should be able to derive similar results.

2. Noise in recurrent learning environments.

   - Is it possible to give sufficient conditions for stability in this setting?

   - How does noise interplay with feedback?

3. Can statements be made about the nature of the relationship between the $\mathcal{C}$'s and E_0, at least for specific learning environments, so that one could optimize, apriori, the test error given a fixed N in order to pick the "best" learning system?