Learning in the Presence of Noise*

Malik Magdon-Ismail
magdon@cco.caltech.edu
Department of Electrical Engineering
California Institute of Technology
136-93 Pasadena, CA, 91125

Alexander Nicholson
zander@foot.caltech.edu
Department of Computer Science
California Institute of Technology
136-93 Pasadena, CA, 91125

Yaser S. Abu-Mostafa
yaser@caltech.edu
Department of Electrical Engineering
and Department of Computer Science
California Institute of Technology
136-93 Pasadena, CA, 91125

Abstract

We report results about the impact of noise on information processing, with application to financial markets. These results quantify the tradeoff between the amount of data and the noise level in the data. They also provide estimates for the performance of a learning system in terms of the noise level. We use these results to derive a method for detecting the change in market volatility from period to period. We successfully apply these results to the four major foreign exchange markets. The results hold for linear as well as non-linear learning models and algorithms, and for different noise models.

Keywords: Learning, Noise, Convergence, Bounds, Test Error, Generalization Error, Model Limitation, Volatility.

1 Introduction

Information processing of financial data entails the extraction of relevant information from overwhelming noise. The levels of noise in financial markets are such that the most one can hope for is 'getting it right' slightly

*To appear in book form in *Intelligent Signal Processing*, IEEE Press. Also appeared as *Financial Markets: Very Noisy Information Processing*, Proceedings of the IEEE, Vol. 86, No. 11, Nov, 1998.

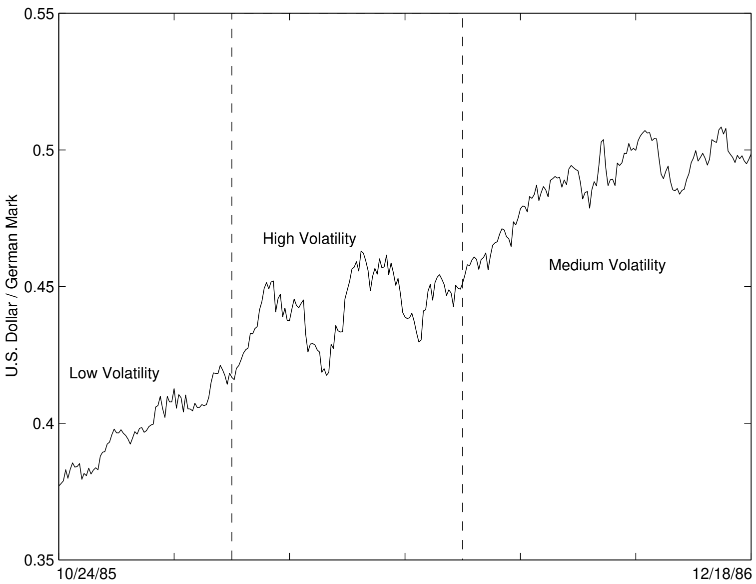

Figure 1: *The price curve for the U.S. Dollar vs. German Mark. The market volatility can change noticeably over the course of time.*

better than 50% of the time [17]. To complicate matters further, one also needs to be reasonably sure that one is not being fooled by a finite set of examples from historical data into believing that the performance is acceptable when it is actually (and disastrously) slightly worse than acceptable.

In addition to being a nuisance that complicates the processing of financial data, noise plays a role as a tradable commodity in its own right. Indeed, market volatility is the basis for a number of financial instruments, such as *options* [6], whose price explicitly depends on the level of volatility in the underlying market. For this reason, it is of economic value to be able to predict the changes in the noise level in financial time series as these changes are reflected in the price changes in tradable instruments. These changes can be significant as one can observe in figure 1 where the U.S. Dollar/German Mark market has undergone extreme changes in volatility.

In spite of the high levels of noise, financial data are among the best application domains for intelligent processing and advanced learning techniques. These data have been recorded very accurately for very long periods of time. They are available on different time scales, and simultaneously available in many different markets. This provides a very rich environment for analysis and experimentation using advanced processing techniques. Moreover, the payoff for even minute, but consistent, improvements in performance is huge.

In this paper, we tackle the question of how information processing is affected by the presence and variability of noise in the data. In doing so, we do not restrict the distribution or the time-varying nature of the noise, nor do we restrict the learning model or learning algorithm that we use. We report results that provide quantitative estimates of the optimal performance that can be achieved in the presence of noise. In financial markets, this provides a benchmark for the target performance given a set of data. We also quantify the tradeoff between the amount of data needed and the level of noise in the data. Our experiments with real foreign exchange data demonstrate that the results are applicable to the case of finite data, the only case of practical interest. They also provide a means of assessing the change in the level of noise in financial data that can be applied to volatility-based financial instruments.

The paper is organized as follows. Section 2 introduces financial time series and section 3 covers the main results about the impact of noise. It should be stressed that in section 3 we analyze a very general learning paradigm, and although the main domain of application here is to the processing of financial time series, the results are applicable to a wide variety of learning problems. These results are tested in the four major foreign exchange markets in section 4. The appendix includes the formal definitions, theorems, and complete proofs of all the results that we report.

2 Financial Time Series Prediction

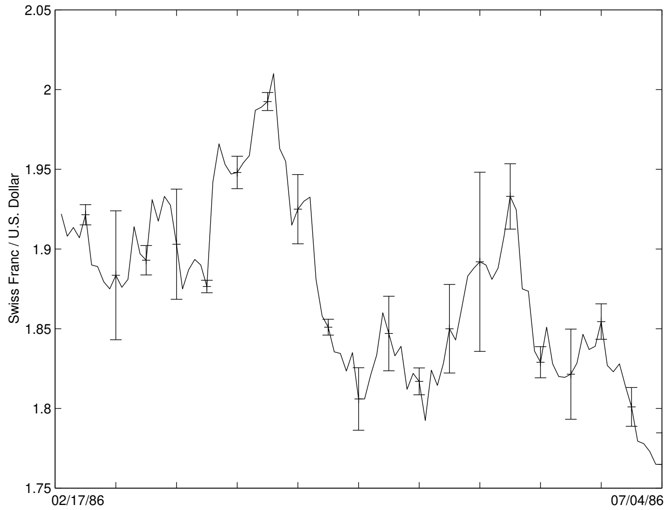

Figure 2: *Financial time series tend to be very volatile. Above is the realization of one such time series from the FX markets. Error bars are used merely to illustrate that each point is the outcome of a noisy event.*

Financial markets present us with data in the form of a time series. We might have the daily, hourly or tick-by-tick stock prices or foreign exchange rates. For financial time series, it is of economic interest to predict the value at some time in the future. Thus, we would like to extract as much information as possible from historical data with the hope of learning the underlying behavior. In general, we can consider the value of a time series $y(t)$ at any time t as a noisy data point $y = f(\mathbf{x}) + \epsilon$. Here f is a deterministic function of a vector $\mathbf{x}(t)$ of market indicators and $\epsilon(t)$ is noise. The task at hand is one of learning $f(\cdot)$ from a finite data set (the history of the series).

In modeling the time variation of a stock price S, the standard Black-Scholes model for pricing options based on volatility [6] assumes the variation to be of the form

$$\mathrm{d}S = \tilde{\mu} S \mathrm{d}t + \tilde{\sigma} S \eta \sqrt{\mathrm{d}t}$$

where $\tilde{\sigma}$ is the market volatility and η has a zero-mean normal distribution with variance 1. Thus, the Black-Scholes model uses only the previous price as the indicator vector $\mathbf{x}$. The price at different times has a deterministic dependence on the past (the $\tilde{\mu}$ term) and a noisy component (the $\tilde{\sigma}$ term). The variance of the noisy component is related to the volatility and need not be constant. The precise relation is given by

$$\sigma_i^2 = \tilde{\sigma}^2 \Delta S_i^2 \tag{1}$$

where i is a time index, and ΔS_i is increment in the stock price from time i to time $i+1$.

Extensive literature already exists on methods for extracting information from noisy time series ([7], [8], [10], [11], [14]). The details of such methods are not our present concern. We are interested in determining how our prediction performance depends on the amount of available data and the variability of the data (which is related to market volatility (1)) – what change in performance are we to expect if this year's market is more volatile than last years market? What change in performance relative to some benchmark are we to expect if the market changed recently and hence we only have few data points to learn from?

Pricing information is available on a variety of time scales, which presents us with a data set size vs. variability tradeoff. We could choose to use the tick-by-tick data because we will then have many data points, but the price we have to pay is that these data points are much noisier. The trade off will depend on how much noisier the tick-by-tick data is, and the details of the learning scheme. Market analysts would like to quantify this tradeoff by how it affects performance.

An estimate of the best performance that we can achieve with a given information extraction scheme might also be economically useful. As well as providing a criterion for selecting between different models, knowing the model limitation could be useful for determining whether even an unlimited amount of data will give a system that is financially worth the risk. This would allow analysts to compare trading strategies based on their model limitation.

It is to be expected that when markets are volatile, the performance of a learning system drops. However, the effects of the noise should become less pronounced with increasing data availability. In the next section, we quantify this intuition.

3 Impact of Noise on Learning

In this section we address the issues raised in section 2 in the context of learning theory. We begin by setting up the learning problem, restate the questions in the learning theory framework, and present theoretical and experimental results.

3.1 The Learning Problem

We assume the standard learning paradigm. The goal is to learn a target function $f : \mathbf{R}^d \to \mathbf{R}$. A training data set $\mathcal{D}_N$ is given, which consists of N input output pairs $\{\mathbf{x}_i, y_i\}_{i=1}^N$. Each $\mathbf{x}_i \in \mathbf{R}^d$ is drawn from some input probability measure $dF(\mathbf{x})$ which we assume to have compact support. Learning entails choosing a hypothesis function g from a collection of candidate functions $\mathcal{H}$. We will assume that the target function f and the candidate functions $g \in \mathcal{H}$ are continuous. The set $\mathcal{H}$ is called the *learning model* because it reflects how we choose to model the target function. The hypothesis function is chosen by a *learning algorithm* $\mathcal{A}$

based on some performance criterion on the data. We assume that $\mathcal{A}$ is a mapping $\mathcal{A} : \mathcal{D}_N \to \mathcal{H}$. A typical learning algorithm might be one that uses gradient descent to select the hypothesis which minimizes the mean squared error on the training set. Given a learning task, we select a particular *learning system*, which takes as input a data set and produces a hypothesis function (see figure 3).

Definition 3.1 *A Learning System $\mathcal{L}$ is a pair $\{\mathcal{A}, \mathcal{H}\}$.*

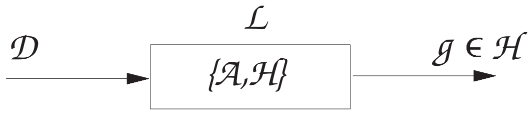

Figure 3: *The learning setup*

Additive noise is present in the training data,

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

We further assume that the noise realizations are independent and zero mean, so

$$\langle \epsilon \mid \mathbf{x} \rangle_\epsilon = \mathbf{0}, \qquad \langle \epsilon\epsilon^T \mid \mathbf{x} \rangle_\epsilon = diag[\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2]$$

(we use $\langle \cdot \rangle$ to denote expectations, $\sigma = [\sigma_1 \sigma_2 \ldots \sigma_N]$, and $diag[\cdot]$ denotes a diagonal matrix). It should be noted that we allow the noise variance to change from one data point to another, which is always the case in financial markets.

Define $g_{\mathcal{D}_N}(\mathbf{x}) \in \mathcal{H}$ as $\mathcal{A}(\mathcal{D}_N)$, the function that was chosen by the algorithm. We define the test error for $g_{\mathcal{D}_N}$ as the expectation of the squared deviation between $g_{\mathcal{D}_N}$ and $f(\mathbf{x})$ taken over the input space. Thus the test error measures the ultimate performance of our system after it has learned from the data. We denote the test error by $E[g_{\mathcal{D}_N}]$.

$$E[g_{\mathcal{D}_N}] = \left\langle (g_{\mathcal{D}_N}(\mathbf{x}) - f(\mathbf{x}))^2 \right\rangle_{\mathbf{x}} \tag{2}$$

We can further define the expected test error, $\mathcal{E}_N(\sigma)$ as the expectation of the test error taken over possible realizations of the noise and the data set.

$$\mathcal{E}_N(\sigma) = \langle E[g_{\mathcal{D}_N}] \rangle_{\epsilon, \mathcal{D}_N} \tag{3}$$

The goal is to minimize $\mathcal{E}_N(\sigma)$. $\mathcal{E}_N(\sigma)$ represents the expected test performance averaged over the choice of training examples. It is related to the "future profit" you expect to make having trained your learning system on the available data. $\mathcal{E}_N(\sigma)$ will depend on the detailed properties of the learning system and target function. It would be a daunting task to tackle the behavior of $\mathcal{E}_N(\sigma)$ in general, but as we shall see, under quite unrestrictive conditions, the changes in $\mathcal{E}_N(\sigma)$ as the noise or data set size change can be quantified. This will be related to the tradeoff in profit when attempting to learn and predict during more volatile stages of the market compared to less volatile stages.

A related quantity of interest is $\mathcal{N}$, the number of data points (with noise added) that are needed to attain a test error comparable to that attainable when N noiseless examples are available.

$$\mathcal{N}(\Delta, \sigma, N) \triangleq \min_{N_1} \{N_1 : \mathcal{E}_{N_1}(\sigma) - \mathcal{E}_N(0) \le \Delta\} \tag{4}$$

$\mathcal{N}(\Delta, \sigma, N)$ is the number of noisy examples that are equivalent to N noiseless examples, and it describes the trade off between numerous, more volatile data, versus fewer and less volatile data. The answers to the questions posed in section 2 lie in the behavior of $\mathcal{E}_N(\sigma)$ and $\mathcal{N}(\Delta, \sigma, N)$. We address these questions analytically next.

3.2 Performance of a Learning System

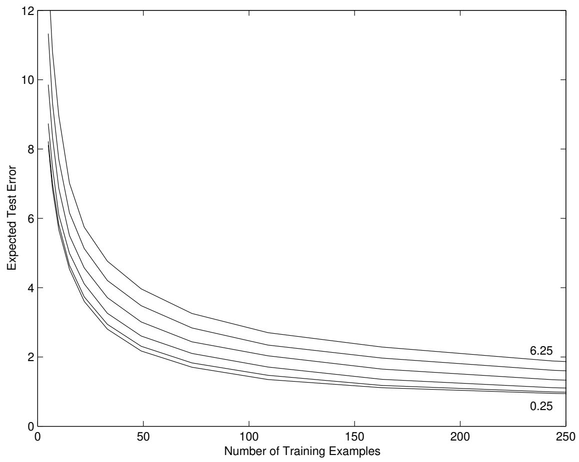

Figure 4: *Experiments illustrating the behavior of the test error as a function of N for various noise levels with variances ranging from 0.25–6.25. A non-linear neural network learning model was used with gradient descent on the squared error. Data was created using a non-linear target function.*

Intuition tells us that noisier data leads to worse test performance. This is because the learning system attempts to fit the noise (i.e. learn a random effect) at the expense of fitting the true underlying dependence. However, the more data we have, the less pronounced the impact of the noise will be. This intuition is illustrated in figure 4. We observe that the higher the noise, the higher the test error, but the curves appear to be getting closer to each other as we use more and more examples for the learning process. We would like to quantify this idea.

In order to be able to do so, we need to restrict ourselves to *stable learning systems*. Stable learning systems possess the two properties "continuity" and "unbiasedness". Continuity ensures that "close" data sets are mapped to "close" functions. For two data sets differing only by the addition of zero-mean noise, unbiasedness requires that at every point, the average value (with respect to the noise) of the functions resulting from the noisy data set is equal to the value of the function resulting from the noiseless data set. (Refer to the appendix for formal definitions.)

These properties are somewhat intuitive, and we note that, for any learning system $\mathcal{L}$, they can be checked directly. We would like our learning procedure to be robust towards small noise fluctuations in the data so we do not consider learning models that may yield discontinuous behavior. The unbiasedness property may seem fragile, especially given the extremely nonlinear nature of a learning algorithm. Nevertheless, we consider it an important and not overly restrictive condition on a learning system. If the noise is small, then the first order change in $\mathcal{A}(D_N)$ should be proportional to the noise parameter, so that the average change is zero with zero-mean noise. Indeed, experiments with neural networks show that learning with gradient descent and conjugate gradient descent on the mean squared error are unbiased with a reasonable noise level. Thus, linear and neural network learning models give learning systems that are stable.

We then have the following theorem.

Theorem 3.2 *Let $\mathcal{L}$ be stable. Then $\forall\epsilon > 0$, $\exists\mathcal{C}_1$, $\mathcal{C}_2$ such that using $\mathcal{L}$, it is at least possible to attain a test error bounded by*

$$\mathcal{E}_N(\sigma) < \mathcal{E}_N(0) + \frac{\overline{\sigma^2}\mathcal{C}_1}{N} + \epsilon + O\left(\frac{1}{N^2}\right) \tag{5}$$

$$\mathcal{E}_N(0) < E_0 + \frac{\mathcal{C}_2}{N} + \epsilon + o\left(\frac{1}{N}\right) \tag{6}$$

where $\lim_{N\to\infty}\mathcal{E}_N(0) = E_0$ and $\overline{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2$. $\mathcal{C}_1, \mathcal{C}_2$ are constants that depend on the input distribution, target function and learning system.

The proof can be found in the appendix (Theorem B.5). Furthermore, in certain cases we can combine (5) and (6) to get

$$\mathcal{E}_N(0) < E_0 + \frac{\mathcal{C}_1\overline{\sigma^2} + \mathcal{C}_2}{N} + o\left(\frac{1}{N}\right) \tag{7}$$

The essential content of the theorem is that the expected test error increases in proportion to $\overline{\sigma^2}$ holding everything else constant, and decreases in proportion to $1/N$ holding everything else constant. The conditions of Theorem 3.2 are quite general and are satisfied by a wide variety of learning models and algorithms. For learning models that are linear $\mathcal{C}_1 = d + 1$. E_0 is the model limitation modulo the learning algorithm when tested on noiseless data. The limiting performance on noisy future data is $E_0 + \overline{\sigma^2}$. One expects that for more complex models, the model limitation (E_0) is lower than for less complex learning models. However, the convergence parameters ($\mathcal{C}_1$, $\mathcal{C}_2$) are expected to be larger for more complex models. Thus, for a given number of data points, there will be an optimal model complexity (eg. number of hidden units for a neural network) minimizing the bound of theorem 3.2. One can compare this tradeoff to the bias-variance tradeoff [19].

Experimentally we observe that the bounds of theorem 3.2 are quite tight even for small N (see figure 5) so combining (5) and (6) we expect the following dependence for $\mathcal{N}(\Delta, \sigma, N)$, the number of noisy examples that are equivalent to N noiseless examples.

$$\mathcal{N}(\Delta, \sigma, N) \sim \frac{\overline{\sigma^2}\mathcal{C}_1 + \mathcal{C}_2}{\frac{\mathcal{C}_2}{N} + \Delta} \tag{8}$$

The results are illustrated in figure 5. Artificial data sets were created from a known target function. Figure 5(a) illustrates the results of fitting a linear model to nonlinear data. Shown is the residual error $\hat{\mathcal{E}}_N(\sigma) = \mathcal{E}_N(\sigma) - \mathcal{E}_N(0)$. The inputs are chosen from $\mathbf{R}^2$, and the dashed lines illustrate that $\hat{\mathcal{E}}_N(\sigma)$ quickly converges to $3\overline{\sigma^2}/N$ as expected from (5). Figure 5(b) shows similar results for a nonlinear learning model. Gradient descent was used to train the three hidden unit neural network model. Ideally we expect this algorithm/model pair to be continuously compatible, and it was empirically shown to be mean preserving. The residual errors

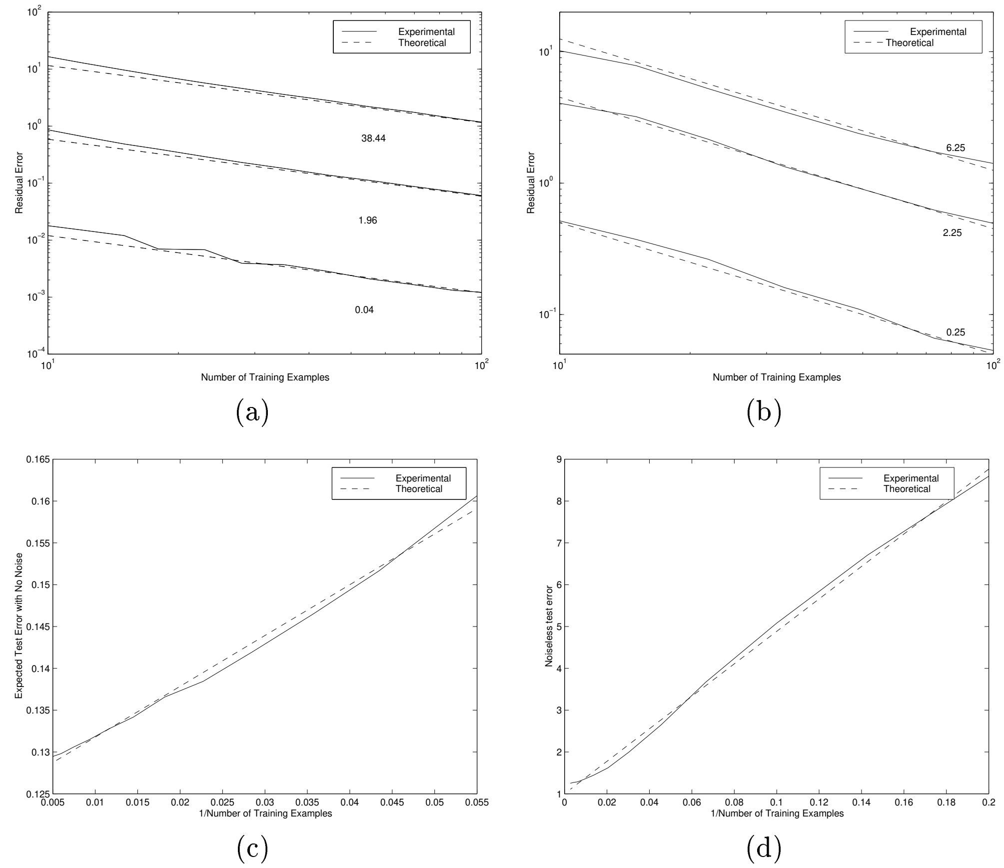

Figure 5: (a)*Nonlinear target function and linear learning model – The residual error is shown for a nonlinear target function trained using a linear model. Gaussian noise with* $\overline{\sigma^2}$ *ranging from 0.04 to 38.44 was added to training sets. The dashed lines show the error level predicted by Theorem B.2.* (b)*Nonlinear target function and nonlinear learning model – The residual error is shown for a nonlinear target function trained using a 2-3-1 network. Gaussian noise with* $\overline{\sigma^2}$ *ranging from 0.25 to 6.25 was added to training sets. The results correspond very closely to those predicted by Theorem 3.2 when* $C_1 = 20$ *(shown with dashed lines).* (c) *The behavior of the expected test error with no noise for the learning scenario in (a). We observe that for even small* N *there is close agreement between the theoretical* $1/N$ *decrease.* (d) *The behaviour of the expected test error with no noise for the non-linear learning scenario in (b). Once again for small* N *we observe the expected behavior.*

very closely follow $20\,\overline{\sigma^2}/N$, showing that we have approximate equality in (5) for $C_1 \sim 20$.[1] Figures 5 (c) and (d) show that $\mathcal{E}_N(0)$ also behaves linearly in $1/N$ for both cases (i.e. it quickly approaches the bound in 6).

3.3 Estimating the Model Limitation

When the learning model is linear, we can show (theorems B.2, B.3) that the expected training error $\mathcal{E}_{tr}(\sigma)$ (the error on the data set) and expected test error approach the same limiting value from opposite sides as $N \to \infty$. Further the rates of convergence to this limiting value are the same. Amari [1] has obtained a similar asymptotic result in the case of nonlinear models when performing gradient descent on the training error. Using the Amari result, we can use our bound on the test error to bound the training error performance. The expected error on a noisy data set, $\mathcal{E}_{test}$ is related to $\mathcal{E}_N(\sigma)$ by $\mathcal{E}_{test}(\sigma) = \mathcal{E}_N(\sigma) + \overline{\sigma^2}$. The experiments demonstrate that the bounds of theorem 3.2 are almost saturated for small N, so, ignoring terms that are $o(1/N)$, and using Amari's result we have

$$E_0 + \overline{\sigma^2} \leq \quad \mathcal{E}_{test}(\sigma) \approx \quad E_0 + \overline{\sigma^2} + \frac{\mathcal{C}_1\overline{\sigma^2} + \mathcal{C}_2}{N} \tag{9}$$

$$E_0 + \overline{\sigma^2} \geq \quad \mathcal{E}_{tr}(\sigma) \approx \quad E_0 + \overline{\sigma^2} - \frac{\mathcal{C}_1\overline{\sigma^2} + \mathcal{C}_2}{N} \tag{10}$$

(in the case of linear learning models we can replace $\mathcal{C}_1$ by $d+1$). From the data set of size N, for $N_1 < N$, we can randomly pick N_1 data points (perform Bootstrapping [18] on the training data). Thus, by varying N_1 in the training phase and observing the error on the training set, we obtain an estimate of the model limitation $E_0 + \overline{\sigma^2}$. This method also immediately furnishes an estimate of $\mathcal{C}_1\overline{\sigma^2} + \mathcal{C}_2$, so we can estimate the parameters that are needed for the bound (7). This is illustrated in the next section where we apply the results presented here to the case of financial time series.

4 Application to Financial Market Forecasting

We can apply the results of section 3 to real financial market data. Figure 4 illustrates the $1/N$ behavior of the residual error $\hat{\mathcal{E}}_N(\sigma)$ for foreign exchange rates.

Daily close exchange rates between 1984 and 1995 were used for the Swiss Franc (CHF), German Mark (DEM), British Pound Sterling (STG) and Japanese Yen (JPY). A linear model was used to learn the future price as a function of the close price of the previous five days.

We performed the following experiments. The last 1000 data points of each time series were held out as a test set. The remaining points were used to create a data set

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

N_1 points were sampled from this set and used to learn. This was repeated to obtain an estimate of the expected test and training error. We show the dependence of the expected test error on the number of training examples in figure 4. Though it is not obvious that the assumptions made to derive the results hold, as with the results on artificial data, the test error seems to not only obey the bound of equation (5), but quickly assumes $1/N$ behavior. Assuming the bounds to be tight for both the test error and training error, we are

[1]This suggests that the condition in Corollary B.6 holds.

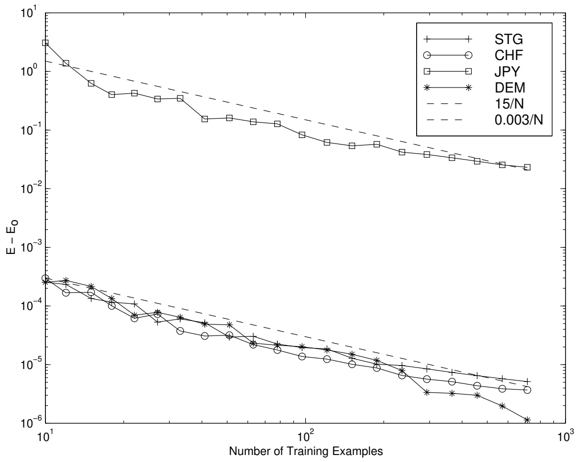

Figure 6: *The dependence of the test error-E_0 on N is depicted for the British Pound (STG), the Swiss Franc (CHF), the Japanese Yen (JPY) and the German Mark (DEM). Also shown are two lines that show $1/N$ behavior. We see that the test error curves follow the theory well.*

able to estimate the best possible performance of the linear model by finding the line best fitting $\mathcal{E}_N(0)$ as a function of $1/N$. Table 4 summarizes these estimates.

We compare the model limitation to that of simply predicting the present value as the next value. We find that this simple strategy virtually attains the model limitation suggesting that today's price completely reflects tomorrow's price – that's the best we can expect to achieve systematically. The results in table 4 are appealing on two accounts. Firstly, assuming that today's price is the best predictor of tomorrow's price, the technique we use to predict the model limitation is performing well. Secondly, because the model limitation estimates are slightly below the error of the simple strategy, we deduce that there is some information that can be extracted from previous prices.

By training on different time periods, we find that the model limitation may change. If we assume the underlying dependence to have remained constant so that E_0 has not changed, then the resulting change can only be due to a change in $\overline{\sigma^2}$ thus providing an estimate of the change in the volatility (since the volatility is related to the change in $\overline{\sigma^2}$ (1)). It appears from table 4 that of the four currencies, the British Pound's volatility seems to have increased while the remaining three markets display decreasing volatility.

We see that the results of section 3 apply to the problem of financial forecasting. Experiment bears out the fact that the answers to the questions posed in section 2 lie in the expressions for $\mathcal{E}_N(\sigma)$ and $\mathcal{N}(\Delta, \sigma, N)$ in equations (5) and (8).

5 Conclusion

The new results in section 2 are represented in theorem 3.2. The experimental results on artificial data amply support the theory. We have shown that the number of noisy examples required for comparable generalization with N noiseless examples increases as $\overline{\sigma^2}$. Explicitly, the main result bounds the test error for a noisy data set

Currency	E_0 Est.	No Change
DEM	0.000499	0.000502
CHF	0.000158	0.000160
STG	0.000134	0.000136
JPY	1.082	1.083

(a)

Currency	E_0 Est.	No Change
DEM	0.000156	0.000152
CHF	0.000148	0.000151
STG	0.000153	0.000157
JPY	0.851	0.867

(b)

Table 1: *Estimate of model limitation and comparison to simple predictor. In (a) we use the training error to estimate* $E_0 + \overline{\sigma^2}$ *and compare to the performance on the training set when we use the simple system: predict no change in price. In (b) we use the test error curve to estimate* E_0*. Only (a) is possible in practice, but both yield very good estimates(if we assume that this simple strategy is close to the best you can do), thus verifying that the results of section 3 can be applied to this learning problem. The change in the estimate from (a) to (b) is due to the fact that the test and training sets are taken from different time intervals, and hence the estimates reflect a change in the market volatility over time.*

by $\mathcal{E}_N(\sigma) \leq E_0 + \frac{\overline{\sigma^2}C_1' + C_2'}{N} + o\left(\frac{1}{N}\right)$. We also obtained a result that bounds the expected test error relative to a benchmark test error (5). Experimentally we showed that this result applies to the non-asymptotic regime – the empirical results show that the bounds hold with almost equality for N as small as 20. Intuitively this is because the non-asymptotic effects that affect $\mathcal{E}_N(\sigma)$ also have a similar effect on $\mathcal{E}_N(0)$.

We began with the goal of answering two questions (initially posed in the context of financial time series): Relative to a benchmark scenario (that of learning with no noise), how does the performance change as the noise and number of examples changes? This dependence is represented by the expression for $\mathcal{E}_N(\sigma)$ above. This expression is a similar result to those derived by Amari [1] and Moody [4]. However the differences are significant. Amari compares the training error when descending on a given error function to the expectation of *that* error when you have finished learning. The learning algorithm is specific but the form of the error function may vary. Moody considers minimization of an error term plus a complexity term and assumes that the input distribution is a sum of delta functions at the training data points. In this paper, we derive a convergence result for the expected squared error without severely restricting the learning algorithm or the input distribution. The results were presented in the context of financial time series analysis, but we note that they are applicable to the general learning problem, independent of most of the details of the learning model and learning algorithm. In particular, we do not require the learning algorithm to minimize a simple training error measure – optimizing a generalized regularized training error (as in [5]) should produce an algorithm that still satisfies the conditions of theorem 3.2.

We provided an estimate of the model limitation which we used to estimate the best possible performance when learning in the FX markets. The results were consistent with the assumption that today's price reflects all the information about tomorrow's price. Using this method for predicting the model limitation, we could detect changes in the market volatility, which is of economic use.

It would be useful to explore the relationship between the constants $(E_0, \mathcal{C}_1, \mathcal{C}_2)$ that parametrize the expected test error dependence.

Acknowledgements

We would like to thank Amir Atiya, Joseph Sill, Zehra Cataltepe and Xubo Song for their helpful comments.

6 Appendix

A Definitions

One expects that if one has "close" data sets $\mathcal{D}_N = \{\mathbf{x}_i, f(\mathbf{x}_i)\}$ and $\mathcal{D}'_N = \{\mathbf{x}_i, f(\mathbf{x}_i) + e(\mathbf{x}_i)\}$ where $e(\mathbf{x}_i)$ is small, then $\mathcal{A}(\mathcal{D}_N)$ should be "close" to $\mathcal{A}(\mathcal{D}'_N)$. For $\mathcal{A}$ to have this property, $\mathcal{H}$ should be able to implement the two "close" functions.[2] We formalize this notion by defining the class of learning systems that are n^{th} order–*continuously compatible* ($\mathcal{CC}_n$) with respect to the probability measure $dF(\mathbf{x})$. We will use the following notation. Let $\mathcal{S}$ be the compact support for $dF(\mathbf{x})$. Let $\mathcal{D} = \{\mathbf{x}_i, y_i\}$, $\mathcal{D}' = \{\mathbf{x}'_i, y'_i\}$ be any two data sets on $\mathcal{S}$ such that for every $\{x_i, y_i\}$ in $\mathcal{D}$ there is a $\{x'_j, y'_j\}$ in $\mathcal{D}'$ such that $\|\mathbf{x}_i - \mathbf{x}'_j\| \le \epsilon_{max}$ and $|y_i - y'_j| \le \epsilon_{max}$ and the same is true for every $\{x'_j, y'_j\}$ in $\mathcal{D}'$, i.e. there is a $\{x_i, y_i\}$ in $\mathcal{D}$ such that $\|\mathbf{x}_i - \mathbf{x}'_j\| \le \epsilon_{max}$ and $|y_i - y'_j| \le \epsilon_{max}$. Let $\mathcal{A}(\mathcal{D}) = g(\mathbf{x})$, $\mathcal{A}(\mathcal{D}') = g(\mathbf{x}) + \eta(\mathbf{x})$.

Definition A.1 *$\mathcal{L}$ is <u>n^{th} order-continuously compatible</u> if $\exists C$ such that*

$$\langle|\eta(\mathbf{x})|^n\rangle_{\mathbf{x}} \le (C\epsilon_{max})^n$$

with probability 1 (i.e. for almost every $\mathcal{D}$). We will write $\mathcal{L} \in \mathcal{CC}_n$.

We would like $\mathcal{A}$ to be "unbiased" in the following sense. If we have a data set $\mathcal{D}$ with $\mathcal{A}(D){=}g_0$ and we add independent, zero mean noise to the targets to get a new data set $\mathcal{D}'$ then we would like $\langle\mathcal{A}(\mathcal{D}')\rangle_{noise} = g_0$, where this average of functions is taken pointwise. This motivates the following definition.

Definition A.2 *Let $\mathcal{D}$ and $\mathcal{D}'$ be two data sets related by $\mathbf{y}' = \mathbf{y} + \epsilon$ where the ϵ_i's are independent and zero mean. Then $\mathcal{L}$ is <u>mean preserving</u> or <u>unbiased</u> if $\langle\mathcal{A}(\mathcal{D}')\rangle_\epsilon = \mathcal{A}(\mathcal{D})$ with probability 1 (i.e. for almost every $\mathcal{D}$).*

Definition A.3 *A learning system $\mathcal{L}$ is <u>stable</u> if it is in $\mathcal{CC}_2$ and it is mean preserving.*

B Proofs of Results

Proposition B.1 *If $\mathcal{L} \in \mathcal{CC}_n$ then $\mathcal{L} \in \mathcal{CC}_m$ for $m = 1 \ldots n$.*

PROOF: By Jensen's inequality, $\langle|f(\mathbf{x})|^a\rangle_{\mathbf{x}} \le \langle|f(\mathbf{x})|\rangle^a_{\mathbf{x}}$ for $0 \le a \le 1$. Letting $f(\mathbf{x}) = \eta(\mathbf{x})^n$ as in Definition A.1 and $a = m/n$ for $m \le n$, the proposition now follows because $\mathcal{L} \in \mathcal{CC}_n$. ■

We use the notation

$$\mathbf{X} = [\,\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots \quad \mathbf{x}_N\,], \qquad \mathbf{y} = \mathbf{f} + \epsilon$$
$$\Sigma \equiv \left\langle\mathbf{x}\mathbf{x}^T\right\rangle_{\mathbf{x}}, \qquad \mathbf{q} = \langle\mathbf{x}f(\mathbf{x})\rangle_{\mathbf{x}}$$

The law of large numbers gives us that $\mathbf{X}\mathbf{X}^T \underset{N\to\infty}{\longrightarrow} N\Sigma$ and $\mathbf{X}\mathbf{f} \underset{N\to\infty}{\longrightarrow} N\langle\mathbf{x}f(\mathbf{x})\rangle_{\mathbf{x}}$. where we assume that the conditions for this to happen are satisfied.

[2] These conditions will often be satisfied in practice.

Theorem B.2 *Let $\mathcal{H}$, the learning model, be the set of linear functions $\mathbf{w}\cdot\mathbf{x}+w_0$ and let the learning algorithm be minimization of the squared error. Then*

$$\mathcal{E}_N(\sigma)=\mathcal{E}_N(0)+\frac{\overline{\sigma^2}(d+1)}{N}+O\left(\frac{1}{N^2}\right) \tag{11}$$

$$\mathcal{E}_N(0)=E_0+\frac{B}{N}+O(\frac{1}{N^{\frac{3}{2}}}) \tag{12}$$

where $E_0=\lim_{N\to\infty}\{\mathcal{E}_N(0)\}$ and B is a constant dependent on the input distribution. It follows that $\mathcal{N}(\Delta,\sigma,N)=\frac{\overline{\sigma^2}(d+1)+B}{\Delta+\frac{B}{N}}+O(\frac{1}{N^{\frac{3}{2}}})$.

Proof:. $g\in\mathcal{L}\Leftrightarrow g(\mathbf{x})=\mathbf{x}^T\mathbf{w}$. The Least Squares estimate of $\mathbf{w}$ is given by

$$\hat{\mathbf{w}}=(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y} \tag{13}$$

from which we calculate

$$\begin{aligned}
\mathcal{E}_N(\sigma) &= \langle\hat{\mathbf{w}}^T\mathbf{x}\mathbf{x}^T\hat{\mathbf{w}}-2\hat{\mathbf{w}}^T\mathbf{x}f(\mathbf{x})+f(\mathbf{x})^2\rangle_{\mathbf{x},\mathbf{X},\epsilon}\\
&= \langle f^2\rangle-2\left\langle\mathbf{f}^T\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\right\rangle_{\mathbf{X}}\mathbf{q}+\left\langle\mathbf{f}^T\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\Sigma(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{f}\right\rangle_{\mathbf{X}}\\
&\quad+\left\langle\epsilon^T\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\Sigma(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\epsilon\right\rangle_{\mathbf{X},\epsilon}\\
&= \mathcal{E}_N(0)+\sum_i\sigma_i^2\left\langle\left[\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\Sigma(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\right]_{ii}\right\rangle_{\mathbf{X}}
\end{aligned}$$

where we have used (1) and $\mathbf{q}=\langle\mathbf{x}f(\mathbf{x})\rangle_{\mathbf{x}}$. By the law of large numbers, we note that $(\mathbf{X}\mathbf{X}^T)^{-1}\underset{N\to\infty}{\longrightarrow}N\Sigma$ and $\mathbf{X}\mathbf{f}\underset{N\to\infty}{\longrightarrow}N\mathbf{q}$, so we write

$$\mathbf{X}\mathbf{X}^T=N\Sigma+\sqrt{N}\mathbf{V}(\mathbf{X}),\qquad\qquad \mathbf{X}\mathbf{f}=N\mathbf{q}+\sqrt{N}\mathbf{a}(\mathbf{X}) \tag{14}$$

where $\langle\mathbf{V}\rangle_{\mathbf{X}}=\langle\mathbf{a}\rangle_{\mathbf{X}}=0$ and $\mathrm{Var}(\mathbf{V})$ and $\mathrm{Var}(\mathbf{a})$ are $O(1)$. Using (14) and the identity $[1+\lambda\mathbf{A}]^{-1}=1-\lambda\mathbf{A}+\lambda^2\mathbf{A}^2+O(\lambda^3)$

$$\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\Sigma(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}=\frac{\mathbf{X}^T\Sigma^{-1}\mathbf{X}}{N^2}-2\frac{\mathbf{X}^T\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{X}}{N^{\frac{5}{2}}}+3\frac{\mathbf{X}^T\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{X}}{N^3}+O\left(\frac{1}{N^{\frac{7}{2}}}\right)$$

From the definition of $\mathbf{X}$ we find from the first term

$$[\langle\mathbf{X}^T\Sigma^{-1}\mathbf{X}\rangle]_{ii}=\left\langle\sum_{k,l}\Sigma_{kl}^{-1}(\mathbf{x}_i)_k(\mathbf{x}_i)_l\right\rangle=[\sum_{k,l}\Sigma_{kl}^{-1}\Sigma_{kl}]_{ii}=1$$

by taking the expectation of the trace of both sides of the equation, the second term can be shown to be of the same order as the third term. So

$$\mathcal{E}_N(\sigma)=\mathcal{E}_N(0)+\frac{\sum_i\sigma_i^2}{N^2}+O(\frac{1}{N^2}) \tag{15}$$

The first part of the theorem now follows. Using similar techniques for $\mathcal{E}_N(0)$, we find

$$\begin{aligned}
\mathcal{E}_N(0) &= \langle f^2\rangle-\mathbf{q}^T\Sigma^{-1}\mathbf{q}+\overbrace{\frac{\langle\mathbf{q}^T\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{q}-\mathbf{a}^T\Sigma^{-1}\mathbf{q}\rangle}{N^{\frac{1}{2}}}}^{0}\\
&\quad+\frac{\langle\mathbf{q}^T\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{q}+\mathbf{a}^T\Sigma^{-1}\mathbf{a}-2\mathbf{a}^T\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{q}\rangle}{N}+O\left(\frac{1}{N^{\frac{3}{2}}}\right)\\
&= E_0+\frac{B}{N}+O\left(\frac{1}{N^{\frac{3}{2}}}\right)
\end{aligned}$$

with $E_0 = \langle f^2 \rangle - \mathbf{q}^T \Sigma^{-1} \mathbf{q}$ and B depending on the input distribution. This gives the N dependence of $\mathcal{E}_N(0)$. Finally we have

$$\mathcal{E}_{\mathcal{N}}(\sigma) - \mathcal{E}_N(0) = \frac{\overline{\sigma^2}(d+1)+B}{\mathcal{N}} - \frac{B}{N} = \Delta$$

yielding the functional dependence $\mathcal{N}(\Delta, \sigma, N)$. ∎

This result can immediately be generalized to the case where the learning model is linear in its parameter space. A similar technique can be used to derive a result on the expected mean squared residual itself which we will call $\mathcal{E}_r(\sigma)$.

Theorem B.3 *Let $\mathcal{H}$, the learning model, be the set of linear functions $\mathbf{w} \cdot \mathbf{x} + w_0$ and let the learning algorithm be minimization of the squared error. Then*

$$\mathcal{E}_r(\sigma) = \mathcal{E}_r(0) + \overline{\sigma^2} - \frac{\overline{\sigma^2}(d+1)}{N} \tag{16}$$

$$\mathcal{E}_r(0) = E_0 - \frac{B}{N} + O(\frac{1}{N^{\frac{3}{2}}}) \tag{17}$$

where E_0 and B are the same constants appearing in Theorem B.2. Thus we find

$$\mathcal{E}_N(\sigma) - \mathcal{E}_r(\sigma) = \frac{2(\overline{\sigma^2}(d+1)+B)}{N} + o(\frac{1}{N^2}) \tag{18}$$

Proof: The residual error is given by

$$\begin{aligned}
\mathcal{E}_r(\sigma) &= \left\langle \frac{1}{N}(\hat{\mathbf{X}}^T\mathbf{w} - \mathbf{y})^2 \right\rangle = \left\langle \frac{1}{N}((\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X} - \mathbf{1})y)^2 \right\rangle \\
&= \frac{\langle \mathbf{f}^T\mathbf{f} \rangle - \left\langle f^T\mathbf{X}(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{f} \right\rangle + \langle \epsilon^T \epsilon \rangle - \left\langle \epsilon^T \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\epsilon \right\rangle}{N} \\
&= \underbrace{\langle f^2 \rangle - \frac{\left\langle \mathbf{f}^T X^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{f} \right\rangle}{N}}_{\mathcal{E}_r(0)} + \overline{\sigma^2} - \frac{\overline{\sigma^2}(d+1)}{N}
\end{aligned}$$

from which the first part of the theorem follows. Using the techniques of Theorem B.2 we find that

$$\begin{aligned}
\mathcal{E}_r(0) &= \langle f^2 \rangle - \mathbf{q}^T\Sigma^{-1}\mathbf{q} \\
&\quad - \frac{\langle \mathbf{q}^T\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{q} + \mathbf{a}^T\Sigma^{-1}\mathbf{a} - 2\mathbf{a}^T\Sigma^{-1}\mathbf{V}\Sigma^{-1}\mathbf{q} \rangle}{N} + O(\frac{1}{N^{\frac{3}{2}}})
\end{aligned}$$

Comparing with Theorem B.2 we have the second part of the theorem. ∎

This result is similar to the results obtained by [1],[4].

We now consider the case of a non-linear learning model. The following proposition shows that $\lim_{N\to\infty} \mathcal{A}(\mathcal{D}_N) = g_\infty$ is well defined pointwise – i.e., $\forall \epsilon > 0$, $\exists M$ such that if $N > M$ then $\max_{\mathbf{x}} |g_\infty - \mathcal{A}(\mathcal{D}_N)| < \epsilon$. This can be skipped if this fact is self evident or if one wishes to assume convergence and one is merely interested in the rate. It is included here purely for technical completeness.

Proposition B.4 *Let $\mathcal{L} \in \mathcal{CC}_2$. Then, the limit $\lim_{N\to\infty} \mathcal{A}(\mathcal{D}_N) = g_\infty$ for noiseless data sets is well defined point wise on sets of non-zero probability – i.e., $\forall \epsilon > 0$, $\exists M$ such that if $N > M$ then $\max_{\mathbf{x}} |g_\infty - \mathcal{A}(\mathcal{D}_N)| < \epsilon$.*

Proof: We will sketch the idea of the proof, the details can be filled in using exactly the same techniques as for the proof to theorem B.9. First we show that for any two infinite data sets, the learned functions are essentially identical. For any infinite data set, as the input support is compact (closed and bounded), any infinitesimal volume of non zero probability has an infinite number of data points. Consider two such data sets. The means of the targets in this small volume will be equal (by the law of large numbers). Because the target function is continuous on this compact support, the means for the two data sets are arbitrarily close to the true values for each data set (this can be attained by letting the the size of the volume be arbitrarily small). By continuous compatibility, these two data sets must both be mapped arbitrarily close to the data set with the means as targets. Therefore they must be mapped arbitrarily close to each other. Thus, we see that $\langle(g_1 - g_2)^2\rangle$ is less than ϵ for arbitrary small ϵ, where the two different data sets drawn from the input distribution are mapped to g_i. So we conclude that $\langle(g_1 - g_2)^2\rangle = 0$, therefore, $g_1 = g_2$ with probability 1. Thus, any two infinite noiseless data sets are mapped to the same function (as the functions are continuous), which we call g_∞.

Finally, consider a data set $\mathcal{D}_N$. For N large enough, this data set can be made arbitrarily close to an infinite data set using the argument above. Let $g_N = \mathcal{A}(\mathcal{D}_N)$. Therefore $\langle(g_N - g_\infty)^2\rangle$ can be made arbitrarily small by choosing N large enough. In other words, $\lim_{N\to\infty}\langle(g_N - g_\infty)^2\rangle = 0$, therefore g_N converges to g_∞ with probability 1. Further, because the functions are continuous and the support is compact, this convergence is uniform. ■

We have just shown that the limit $\mathcal{A}(\mathcal{D}_N)$ exists as $N \to \infty$. Thus, with noiseless data sets, we have convergence for stable learning systems. We now consider both the rate of convergence and what happens when noise is added.

Theorem B.5 *Let $\mathcal{L}$ be stable. Let the target function f be continuous. Let the probability measure on the input space have compact support $\mathcal{X}$. Then $\forall\epsilon > 0$, $\exists C_1 > 0$ such that using $\mathcal{L}$, it is at least possible to attain a test error bounded by*

$$\mathcal{E}_N(\sigma) < \mathcal{E}_N(0) + \frac{\overline{\sigma^2}C_1}{N} + \epsilon + O\left(\frac{1}{N^2}\right) \tag{19}$$

Corollary B.6 *If $C_1 \leq C_1' \ \forall\epsilon$ then*

$$\mathcal{E}_N(\sigma) \leq \mathcal{E}_N(0) + \frac{\overline{\sigma^2}C_1'}{N} + \left(\frac{1}{N^2}\right) \tag{20}$$

Corollary B.7 $\lim\limits_{N\to\infty} \mathcal{E}_N(\sigma) = \lim\limits_{N\to\infty} \mathcal{E}_N(0)$*, independent of σ.*

Proof: By rescaling, we can assume that the input space $\mathcal{X} \subseteq \mathcal{S} = [0,1]^d$. f is continuous, so it is uniformly continuous on the compact set $\mathcal{S}$. Therefore, $\exists\delta_1$ such that

$$\mid \mathbf{x} - \mathbf{x}' \mid < \delta_1 \Rightarrow \mid f(\mathbf{x}) - f(\mathbf{x}') \mid < \delta_2$$

Divide $[0,1]$ into intervals of size $\delta_1/\sqrt{d}$. Thus we divide $\mathcal{S}$ into $\left(\sqrt{d}/\delta_1\right)^d$ cubes. Let $C_\mathbf{i} \equiv C_{i_1,i_2\ldots i_d}$ define the cube with lowest coordinates $\mathbf{i}$. Let $N_\mathbf{i}$ be the number of data points in $C_\mathbf{i}$, and let $\mu_\mathbf{i} = \frac{1}{N_\mathbf{i}}\sum_{x_j\in C_\mathbf{i}} y_j$. Let $P_\mathbf{i} = Pr\{x \in C_\mathbf{i}\}$. We only need consider regions where $P_\mathbf{i} > 0$, as regions with $P_\mathbf{i} = 0$ are don't care regions. The following Lemma is easily obtained by noting that for $\mathbf{x}, \mathbf{x}' \in C_\mathbf{i}$, $|f(\mathbf{x}) - f(\mathbf{x}')| \leq \delta_2$.

Lemma B.8 *Let $\mathbf{x} \in C_\mathbf{i}$*

$$\left|\frac{1}{N_\mathbf{i}}\sum_{\mathbf{x}_j\in C_\mathbf{i}} y_j - f(\mathbf{x})\right| = \mid \mu_\mathbf{i} - f(\mathbf{x}) \mid \leq \delta_2 + \frac{\mid \sum\limits_{x_k\in C_i} \epsilon_k \mid}{N_\mathbf{i}}$$

Construct a new data set by replacing all the y's in $C_\mathbf{i}$ by $\mu_\mathbf{i}$. i.e., with no noise, the targets would be $f(\mathbf{x}_j)$ and with noise they are $\mu_\mathbf{i}$. $\forall\mathbf{x}_j \in C_\mathbf{i}$,

$$\mu_i = f(\mathbf{x}_j) + \underbrace{\sum_{\mathbf{x}_k\neq\mathbf{x}_j} \frac{f(\mathbf{x}_k) - f(\mathbf{x}_j)}{N_\mathbf{i}}}_{|\cdot| < \delta_2} + \frac{\sum\limits_{x_k\in C_i} \epsilon_k}{N_\mathbf{i}} = f(\mathbf{x}_j) + \eta_j + \xi_j$$

where $\eta_j = \sum_{\mathbf{x}_k \neq \mathbf{x}_j} \frac{f(\mathbf{x}_k)-f(\mathbf{x}_j)}{N_{\mathbf{i}}}$ and $\xi_j = \sum_{x_k \in C_i} \epsilon_k / N_{\mathbf{i}}$. We have that $\langle \eta_j \rangle_{\mathcal{D}_N} = 0$ and $\langle \xi_j \rangle_\epsilon = 0$. Let $\mathcal{A}$ map the noiseless data set to $g_0 \in \mathcal{H}$ and this noisy version of the data set to $g = g_0 + \eta$. So for the test error we have

$$\mathcal{E}_N(\sigma) = \left\langle (f-g)^2 \right\rangle_{\mathcal{D}_N,\mathbf{x},\epsilon} = \underbrace{\left\langle (f-g_0)^2 \right\rangle_{\mathcal{D}_N,\mathbf{x},\epsilon}}_{\mathcal{E}_N(0)} + \underbrace{2\left\langle (f-g_0)(g_0-g) \right\rangle_{\mathcal{D}_N,\mathbf{x},\epsilon}}_{T_1} + \underbrace{\left\langle (g_0-g)^2 \right\rangle_{\mathcal{D}_N,\mathbf{x},\epsilon}}_{T_2}$$

We now examine T_1 and T_2.

$$\begin{aligned}
|T_1| &= \left| \left\langle (f-g_0) \langle g_0 - g \rangle_\epsilon \right\rangle_{\mathcal{D}_N,\mathbf{x}} \right| \\
&\overset{(a)}{=} \left| \left\langle (f-g_0)(\mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)\}) - \mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)+\eta_k\})) \right\rangle_{\mathcal{D}_N,\mathbf{x}} \right| \\
&\leq \left| \left\langle f \left\langle (\mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)\}) - \mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)+\eta_k\})) \right\rangle_{\mathcal{D}_N} \right\rangle_{\mathbf{x}} \right| \\
&\quad + \left| \langle g_0 (\mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)\}) - \mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)+\eta_k\})) \rangle_{\mathcal{D}_N \mathbf{x}} \right| \\
&\overset{(b)}{\leq} \left\langle \max_{\mathbf{x}} |\, g_0| \left\langle |\mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)\}) - \mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)+\eta_k\})| \right\rangle_{\mathbf{x}} \right\rangle_{\mathcal{D}_N} \\
&\overset{(c)}{\leq} C\delta_2 \left\langle \max_{\mathbf{x}} |\, g_0| \right\rangle_{\mathcal{D}_N} \\
&\overset{(d)}{\leq} c_1 \delta_2
\end{aligned}$$

where (a) and (b) follow from the mean preserving assumption. (c) from continuous compatibility and (d) because we assume the limit g_∞ to exist pointwise. Similarly, for T_2 we get

$$\begin{aligned}
|T_2| &= \left| \left\langle (\mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)\}) - \mathcal{A}(\{\mathbf{x}_k, f(\mathbf{x}_k)+\eta_k+\xi_k\}))^2 \right\rangle_{\mathcal{D}_N,\mathbf{x},\epsilon} \right| \\
&\overset{(a)}{\leq} 2C^2 \left(\delta_2^2 + \left\langle \sum_{\mathbf{i}} \frac{\sum_{\mathbf{x}_j,\mathbf{x}_k \in C_{\mathbf{i}}} \epsilon_j \epsilon_k}{N_{\mathbf{i}}^2} \right\rangle_{\mathcal{D}_N,\epsilon} \right) \\
&= 2C^2 \left(\delta_2^2 + \left\langle \sum_{\mathbf{i}} \frac{\bar{\sigma}_i^2}{N_{\mathbf{i}}} \right\rangle_{\mathcal{D}_N,\epsilon} \right) \\
&\overset{(b)}{=} 2C^2 \left(\delta_2^2 + \overline{\sigma^2} \left\langle \sum_{\mathbf{i}} \frac{1}{N_{\mathbf{i}}} \right\rangle_{\mathcal{D}_N} \right) \\
&= 2C^2\delta_2^2 + 2C^2\overline{\sigma^2} \sum_{\mathbf{i}} \underbrace{\sum_{n=1}^{N} \frac{1}{n} \binom{N}{n} P_{\mathbf{i}}^n (1-P_{\mathbf{i}})^{N-n}}_{\frac{1}{NP_{\mathbf{i}}} + O\left(\frac{1}{N^2}\right)} \\
&= 2C^2\delta_2^2 + \frac{\overline{\sigma^2}}{N} \underbrace{2C^2 \sum_{\mathbf{i}} \frac{1}{P_{\mathbf{i}}}}_{C_1} + O\left(\frac{1}{N^2}\right)
\end{aligned}$$

(a) follows from the continuous compatibility assumption. (b) follows because the noise is chosen independently of the inputs. Choosing δ_1 such that $c_1\delta_2 + 2C^2\delta_2^2 < \epsilon$ we have

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

■

We note that it is easy to extend these theorems to the case where the noise variances are drawn from some distribution. By taking the expectation over that distribution, the same result with $\overline{\sigma^2}$ being the expected value of the variance parameter is obtained. Note also that the preceding proof is by no means suggesting a method to calculate C_1. It is simply a means to show its existence. Often, especially when the input distribution is bounded, Corollary B.6 will hold, and it might be possible to estimate these constants experimentally.

One might wonder what would happen if the mean preserving assumption is violated. We note that the only place where this is used is in the evaluation of T_1. Continuity could still be used however with the difference being that a term of order $\epsilon/\sqrt{N}$ would remain. in other words, one would have $\mathcal{E}_N(\sigma) \leq \mathcal{E}_N(0) + C''\sigma/\sqrt{N}$ + higher order. So if we do not have the mean preserving property then these methods do not guarantee $1/N$ convergence of the test error. Using identical methods, one can, however, get the following result using the continuity property alone: $\langle|f-g|\rangle \leq \langle|f-g_0|\rangle + \frac{C'''\sigma}{\sqrt{N}}$. This is very similar to Theorem B.5 where one measures test error by the expectation of the magnitude difference as opposed to the squared difference.

We now derive a theorem on the dependence of $\mathcal{E}_N(0)$.

Theorem B.9 *Let $\mathcal{L}$ be stable. Let the target function f be continuous. Let the probability measure on the input space have compact support. Then $\forall\epsilon > 0$, $\exists C_2 > 0$ such that using $\mathcal{L}$ it is at least possible to attain a noiseless test error bounded by*

$$\mathcal{E}_N(0) < E_0 + \frac{C_2}{N} + \epsilon + o\left(\frac{1}{N}\right) \tag{21}$$

Corollary B.10 *If $C_2 \leq C_2'$ $\forall\epsilon$ then $\mathcal{E}_N(0) \leq E_0 + \frac{C_2'}{N} + o\left(\frac{1}{N}\right)$ where $E_0 = \lim_{N\to\infty} \mathcal{E}_N(0)$*

Before we proceed to the proof of the theorem, the following lemma is needed.

Lemma B.11 *Let N balls be independently be distributed into r cells according to the probabilities $p_1 \ldots p_r$. Then for every $m > 0$,$\exists A_m$ such that the probability, q, that at least one cell is empty is bounded by*

$$q \leq \frac{A_m}{N^m}$$

PROOF:

$$\begin{aligned} q &= Pr[\cup\ cell_i \text{ is empty}] \leq \sum_i Pr[cell_i \text{ is empty}] = \sum_i (1-p_i)^N \\ &\leq r(1-\min_i p_i)^N \leq \frac{A_m}{N^m} \end{aligned}$$

choosing $A_m \geq r(-m/ln(a))^m$, where $a = 1 - \min_i p_i$. ■

PROOF OF THEOREM B.9

Let $\mathcal{X}$, $\mathcal{S}$, δ_1, δ_2, $C_\mathbf{i}$, $P_\mathbf{i}, N_\mathbf{i}$ be as in the proof of Theorem B.2. We only consider those cubes with $P_\mathbf{i} > 0$.

Suppose that we have an infinite noiseless data set, $\mathcal{D}_\infty$. For all $\mathbf{i}$, let $\bar{y}_\mathbf{i} = \langle f(\mathbf{x})\rangle_{\mathbf{x}|\mathbf{x}\in C_\mathbf{i}}$ and let $\mu_\mathbf{i} = \frac{1}{N_\mathbf{i}}\sum_{\mathbf{x}_j\in C_\mathbf{i}} y_j$ if $C_\mathbf{i}$ is non-empty else, $\mu_\mathbf{i} = 0$. Construct two data sets from the infinite one, $\mathcal{D}_1$ and $\mathcal{D}_2$, by replacing all the y's in $C_\mathbf{i}$ by $\bar{y}_\mathbf{i}$, and $\mu_\mathbf{i}$ respectively. $\mathcal{D}_1$ does not depend on $\mathcal{D}_N$ and $\mathcal{D}_2$ can be obtained from $\mathcal{D}_N$. $\mathcal{D}_\infty$ and $\mathcal{D}_1$ are close data sets because for $\mathbf{x} \in C_\mathbf{i}$,

$$\left|f(\mathbf{x}) - \langle f(\mathbf{y})\rangle_{\mathbf{y}|\mathbf{y}\in C_\mathbf{i}}\right| = \left|\langle f(\mathbf{x}) - f(\mathbf{y})\rangle_{\mathbf{y}|\mathbf{y}\in C_\mathbf{i}}\right| \leq \langle|f(\mathbf{x}) - f(\mathbf{y})|\rangle_{\mathbf{y}|\mathbf{y}\in C_\mathbf{i}} \leq \delta_2$$

Therefore by continuous compatibility, $\left\langle(g_\infty - g_1)^2\right\rangle \leq C^2\delta_2^2$. Define $\epsilon_\mathbf{i}$ by $\mu_\mathbf{i} = \bar{y}_\mathbf{i} + \epsilon_\mathbf{i}$. Then $\langle\epsilon_\mathbf{i}\rangle_{\mathcal{D}_N} = 0$ for all non-empty $C_\mathbf{i}$. Let g_∞, g_1 and g_2 be $\mathcal{A}(\mathcal{D}_\infty)$, $\mathcal{A}(\mathcal{D}_1)$ and $\mathcal{A}(\mathcal{D}_2)$ respectively. Since we can construct $\mathcal{D}_2$, using $\mathcal{L}$ we can at least obtain a test error given by

$$\begin{aligned}\mathcal{E}_N(0) \;\le\;& \overbrace{\left\langle (f-g_\infty)^2\right\rangle}^{E_0} + \overbrace{\left\langle (g_\infty - g_1)^2\right\rangle}^{\le C^2\delta_2^2} + \left\langle (g_1-g_2)^2\right\rangle \\ &+2\underbrace{|\left\langle (f-g_\infty)(g_\infty-g_1)\right\rangle|}_{\substack{\le |f-g_\infty|_{max} C\delta_2 \\ \text{(by cont. comp.)}}} + 2|\left\langle (f-g_\infty)(g_1-g_2)\right\rangle| + 2|\left\langle (g_\infty-g_1)(g_1-g_2)\right\rangle|\end{aligned}$$

By the mean preserving property, $\langle (g_1-g_2)\rangle_{\mathcal{D}_N} = 0$. Therefore,

$$|\left\langle (f-g_\infty)(g_1-g_2)\right\rangle| = |\left\langle (f-g_\infty)\left\langle (g_1-g_2)\right\rangle_{\mathcal{D}_N}\right\rangle_{\mathbf{x}}| = 0$$

Similar reasoning shows that $2|\left\langle (g_\infty-g_1)(g_1-g_2)\right\rangle| = 0$. Let Q be the probability that at least one cell is empty. For the final term we have

$$\begin{aligned}\left\langle (g_1-g_2)^2\right\rangle &\overset{(a)}{\le} (1-Q)C^2\left\langle \sum_{\mathbf{i}} \epsilon_{\mathbf{i}}^2\right\rangle_{\mathcal{D}_N|\forall \mathbf{i}, N_{\mathbf{i}}>0} + 4QC^2|f|^2_{max} \\ &\le C^2\left\langle\left\langle \sum_{\mathbf{i}} \epsilon_{\mathbf{i}}^2\right\rangle_{y_{\mathbf{i}}|N_{\mathbf{i}}>0}\right\rangle_{N_{\mathbf{i}}>0} + c_1 Q \\ &\le C^2\left\langle \sum_{\mathbf{i}} \frac{\sigma_{\mathbf{i}}^2}{N_{\mathbf{i}}}\right\rangle_{N_{\mathbf{i}}>0} + c_1 Q \\ &\overset{(b)}{\le} \frac{1}{N}\underbrace{C^2\sum_{\mathbf{i}}\frac{\sigma_{\mathbf{i}}^2}{P_{\mathbf{i}}}}_{C_2} x + o(\frac{1}{N})\end{aligned}$$

where $\sigma_{\mathbf{i}}^2 = Var(y_{\mathbf{i}}|y \in C_{\mathbf{i}})$. (a) follows by continuous compatibility because with probability $1-Q$ the data sets are at most $\epsilon_{\mathbf{i}}$ apart and $\sum_{\mathbf{i}} \epsilon_{\mathbf{i}}^2 \ge \max_i \epsilon_{\mathbf{i}}^2$, and with probability Q they are at most $2|f|_{max}$ apart. (b) follows because $\langle 1/N_{\mathbf{i}}\rangle = 1/(NP_{\mathbf{i}}) + o(1/N)$ and Lemma B.11 can be used to yield $Q = o(1/N)$. Finally we have

$$\mathcal{E}_N(0) \le E_0 + C^2\delta_2^2 + 2|f-g_\infty|_{max}\delta_2 + \frac{C_2}{N} + o(\frac{1}{N})$$

Choosing δ_2 small enough, we have the theorem because $|f-g_\infty|$ is bounded on the compact support $\mathcal{X}$. ■

References

[1] Murata, N., Yoshizawa, S. and Amari, S., "Learning Curves, Model Selection and Complexity of Neural Networks", *Advances in Neural Information Processing Systems* 5, 1993, pp 607–614.

[2] Amari, S., Murata, N.,Müller, K. R. and Yang, H., "Asymptotic Statistical Theory of Overtraining and Cross Validation", *IEEE Transactions on Neural Networks*, Vol 8, No. 5, pp 985–996 1997.

[3] Amari, S., Murata, N., "Statistical Theory of Learning Curves under Entropic Loss Criterion.", *Neural Computation*, 5, pp 140–153, 1992.

[4] Moody, J. "The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems", *Advances in Neural Information Processing Systems* 4, 1992, pp. 847–854.

[5] Plaut, D., Nowlan, S. and Hinton, G. (1986), "Experiments on Learning by Backpropagation", *Technical Report CMU-CS-86-126*, Carnegie Mellon University.

[6] Black, F. and Scholes, M. S., "The Pricing of Options and Corporate Liabilities", *Journal of Political Economy* 3, 1973, pp. 637-654.

[7] Montgomery, D., Johnson, L. and J. Gardiner, *Forecasting and Time Series Analysis*, New York, McGraw-Hill, Inc., 1990.

[8] Abu-Mostafa, Y. S. and Atiya, A. F., "Introduction to Financial Forecasting", *Applied Intelligence*, 6, 1996, pp 205–213.

[9] Shiller, R. J., *Market Volatility*, Cambridge, MA, MIT Press, 1993.

[10] Trippi, R. R. and Turban, E., *Neural Networks in Finance and Investing*, Chicago, Probos Publishing Company, 1993.

[11] White, H., "Economic Prediction Using Neural Networks: The case of IBM Daily Returns", *Proceedings of the IEEE International Conference on Neural Networks*, 2, 1988, pp 451-458.

[12] Krogh, A. and Hertz, J. A., "Generalization in a Linear Perceptron in the Presence of Noise", *Journal of Physics A* 25, 1992, pp 1135–1147.

[13] Krogh, A., "Learning with Noise in a Linear Perceptron", *Journal of Physics A* 25, 1992, pp 1119–1133.

[14] Abu-Mostafa, Y. S., "Learning from Hints", Journal of Complexity 10, 1994, pp 165–178.

[15] Vapnik, V. N. and Chervonenkis, A., "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities", *Theory Prob. Appl.* 16, 1971, pp 264–280.

[16] Leich, G. and Tanner, J. E., "Economic Forecast Evaluation: Profit versus the Conventional Error Measures", *American Economic Review*, 81, 1991, pp 580–590.

[17] Malkiel, B., *A Random Walk Down Wall Street*, New York, W. W. Norton & Co., 1985.

[18] Shao, J. and Tu, D., *The Jackknife and the Bootstrap*, New York, Springer-Verlag, 1996.

[19] Geman, S. and Bienenstock, E., "Neural Networks and the Bias Variance Dilemma", *Neural Computation*, 4, 1992, pp 1–58