Estimating Model Limitation in Financial Markets

Malik Magdon-Ismail[1], Alexander Nicholson[2] and Yaser Abu-Mostafa[3]

[1] malik@work.caltech.edu
[2] zander@work.caltech.edu
[3] yaser@caltech.edu
Learning Systems Group, California Institute of Technology
136-93 Caltech, Pasadena, CA, USA, 91125

Abstract. We introduce bounds on the generalization ability when learning with noisy data. These results quantify the trade-off between the amount of data and the noise level in the data. Our results can be used to derive a method for estimating the model limitation for a given learning problem. Changes in model imitation can then be used to detect a change in market volatility. Our results apply to linear as well as nonlinear models and algorithms, and to different noise models. We successfully apply our methods to the four major foreign exchange markets.

1 Introduction

Learning from financial data entails the extraction of relevant information from overwhelming noise. Financial markets are dynamic systems so the noise parameters may fluctuate with time. 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* [1], 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 this paper we apply results from learning theory to the task of financial time series prediction. We begin by addressing the problem of learning from noisy data and how learning performance is affected by the presence and variability of noise in the data. We do not restrict the distribution or the time-varying nature of the noise, nor do we place severe restrictions on the learning model or learning algorithm that we use.

Our results 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 data set. We also quantify the trade-off between the noise level and the number of data points used. Our experiments with real foreign exchange data demonstrate that the results are

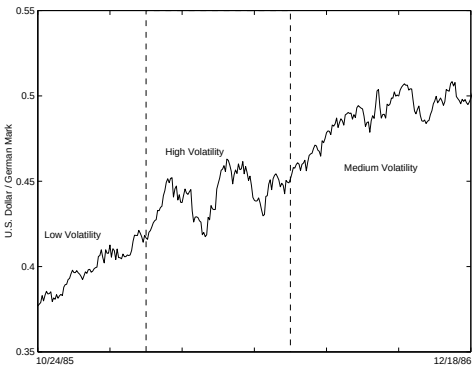


Fig. 1. The price curve for the U.S. Dollar vs. German Mark, illustrating changes in volatility over time.

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.

Section 2 outlines the learning problem and introduces the notation used in the paper. In section 3 we introduce convergence results for stable learning systems, and provide bounds for the test error. These results are then tested in the four major foreign exchange markets in section 4.

2 The Learning Scenario

We assume the standard learning paradigm. The goal is to learn a target function $f : \mathbf{R}^d \to \mathbf{R}$. The training data set, $\mathcal{D}_N$, 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. We will assume that the target function f and the candidate functions $g \in \mathcal{H}$ are continuous. 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.

Let $g_{\mathcal{D}_N}(\mathbf{x}) \in \mathcal{H}$ be $\mathcal{A}(\mathcal{D}_N)$, the function that was chosen by the algorithm. The test error we will be interested in is the expectation of the squared deviation between $g_{\mathcal{D}_N}$ and $f(\mathbf{x})$ taken over the input space. We will denote the test error by $E[g_{\mathcal{D}_N}]$.

$$E[g_{\mathcal{D}_N}] = \left\langle (g_{\mathcal{D}_N}(\mathbf{x}) - f(\mathbf{x}))^2 \right\rangle_{\mathbf{x}} \tag{1}$$

The expected test error, $\mathcal{E}_N(\sigma)$, is then given by

$$\mathcal{E}_N(\sigma) = \langle E[g_{\mathcal{D}_N}] \rangle_{\epsilon, \mathcal{D}_N} \tag{2}$$

The goal is to minimize $\mathcal{E}_N(\sigma)$. $\mathcal{E}_N(\sigma)$ 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.

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

$$\mathcal{N}(\Delta, \sigma, N) \stackrel{\Delta}{=} \min_{N_1} \{ N_1 : \mathcal{E}_{N_1}(\sigma) - \mathcal{E}_N(0) \leq \Delta \} \tag{3}$$

$\mathcal{N}(\Delta, \sigma, N)$ 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. We would like to analyze the behavior of $\mathcal{E}_N(\sigma)$ and $\mathcal{N}(\Delta, \sigma, N)$. We address these questions analytically in section 3, restricting our analysis to the class of *stable* learning systems. These systems have the intuitive properties of "unbiasedness" and "continuity." These concepts are formally defined and some commonly used learning systems are experimentally shown to be stable in [2].

3 Learning System Performance

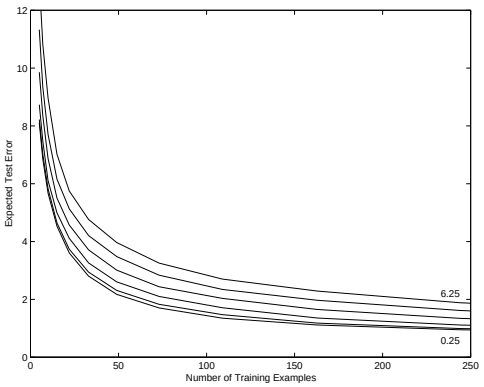
Intuition tells us that noisier data leads to worse test performance. This is because the learning system attempts to fit the noise (i.e. to learn a random effect) at the expense of fitting the true underlying dependence. However, the more data we have, the less pronounced the impact of the noise will be. This intuition is illustrated in figure 2. We observe that the higher the noise, the higher the test error. However, the curves appear to be approaching each other as we use more examples for the learning process. The following theorem quantifies this intuition.

Theorem 3.1 *Let $\mathcal{L}$ be stable. Then, for any $\epsilon > 0$, $\exists \mathcal{C}_1$ such that using $\mathcal{L}$, it is at least possible to attain a test error bounded by*

$$\mathcal{E}_N(\sigma) < \mathcal{E}_N(0) + \frac{\overline{\sigma^2}\mathcal{C}_1}{N} + \epsilon + O\left(\frac{1}{N^2}\right) \tag{4}$$

$$\mathcal{E}_N(0) < E_0 + \frac{\mathcal{C}_2}{N} + \epsilon + o\left(\frac{1}{N}\right) \tag{5}$$

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 generally depend on the input distribution, target function, learning system and possibly ϵ.



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.

Fig. 2. Experiments illustrating the behavior of the test error as a function of N and σ^2.

For a detailed proof of the theorem see [2].

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. When $N \to \infty$, the performance approaches the best attainable independent of the noise level. The conditions of theorem 3.1 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 $E_0 + \overline{\sigma^2}$.

Experimentally we observe that the bounds of theorem 3.1 are quite tight even for small N so combining (4) and (5) we expect the following dependence for $\mathcal{N}(\Delta, \sigma, N)$, the number of noisy examples that are equivalent to N noiseless examples.

$$\mathcal{N}(\Delta, \sigma, N) \sim \frac{\overline{\sigma^2}\mathcal{C}_1 + \mathcal{C}_2}{\frac{\mathcal{C}_2}{N} + \Delta} \tag{6}$$

The constants $\mathcal{C}_1, \mathcal{C}_2,\ E_0$ in theorem 3.1 control the trade-off that affects the sensitivity to noise and convergence rate versus *bias*. Simpler models will have a high value for E_0 but $\mathcal{C}_1$ and $\mathcal{C}_2$ will be small. More complex learning models will have a lower model limitation E_0 but higher convergence parameters $\mathcal{C}_1$ and B. For a given number of data points, there will be some "optimal model complexity."

3.1 Estimating the Model Limitation

When the learning model is linear, we can show 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 $N \to \infty$ ([2]). Further the rates of convergence to this limiting value are the same. In [3], Murata et al. obtained a similar asymptotic result in the case of nonlinear models when performing gradient descent on the training error. Using the Murata 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 3.1 are almost saturated for small N, so, ignoring terms that are $o(1/N)$, and using Murata's result, we have

$$E_0 + \overline{\sigma^2} \leq \mathcal{E}_{test}(\sigma) \leq E_0 + \overline{\sigma^2} + \frac{\mathcal{C}_1\overline{\sigma^2} + \mathcal{C}_2}{N} \tag{7}$$

$$E_0 + \overline{\sigma^2} \geq \ \mathcal{E}_{tr}(\sigma) \geq \ E_0 + \overline{\sigma^2} - \frac{\mathcal{C}_1\overline{\sigma^2} + \mathcal{C}_2}{N} \tag{8}$$

(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 [4] on the training data). By varying N_1 in the training phase and observing the error on the training set, we can obtain an estimate of the model limitation $E_0 + \overline{\sigma^2}$ and an estimate of $\mathcal{C}_1\overline{\sigma^2} + \mathcal{C}_2$. Thus we can estimate the parameters that are needed for the bounds (7) by fitting (8) to the observed dependence of the training error on N_1. In the next section we apply these results to the case of financial time series.

4 Application to Financial Market Forecasting

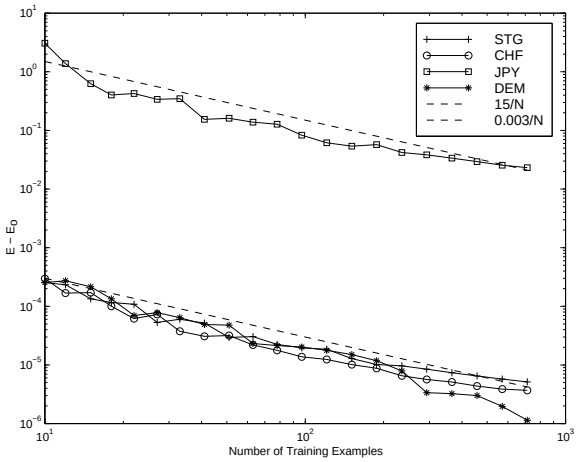
Financial markets present us with data in the form of a time series. In general, we can consider the value of a time series $y(t)$ at any time t as a noisy data point $y = f(\mathbf{x}) + \epsilon$. Here f is a deterministic function of a vector $\mathbf{x}(t)$ of market indicators and $\epsilon(t)$ is noise. The task at hand is one of learning $f(\cdot)$ from a finite data set (the history of the series). The variance $\sigma^2_{\epsilon(t)}$ is related to the volatility ($\tilde{\sigma}$) according to the Black-Scholes Formulation ([1]).

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) – 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? These quantities can be obtained from $\mathcal{E}_N(\sigma)$. $\mathcal{E}_N(\sigma)$ is related to the "future profit" you expect to make having trained your learning system on the available data. Changes in $\mathcal{E}_N(\sigma)$ will be related to the trade-off in profit when attempting to learn and predict during more volatile stages of the market compared to less volatile stages.

Pricing information is available on a variety of time scales, which presents us with a data set size vs. variability trade-off. 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 trade-off by how it affects performance. This trade-off is captured by $\mathcal{N}(\Delta, \sigma, N)$.

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.

Our experimental simulations suggest that we can apply the results of section 2 to real financial market data. Figure 3 illustrates the $1/N$ behavior of the residual error $\hat{\mathcal{E}}_N(\sigma)$ for foreign exchange rates. Daily close exchange rates



The results are 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.

Fig. 3. The dependence of the test error-E_0 on N in some currency markets.

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 3. 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 (4), but quickly assumes $1/N$ behavior. Assuming the bounds to be tight for both the test error and training error, we are 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 1 summarizes these estimates.

Currency	$E_0 + \sigma^2$ 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	$E_0 + \sigma^2$ 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 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 2 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 (assuming E_0 remained constant).

Table 1. Estimate of model limitation and comparison to simple predictor.

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 1 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 (table 1 (a)). That today's price is the best predictor of tomorrow's price is illustrated by table 1 (b) where the $E_0 + \sigma^2$ estimate is the true model limitation estimated using the test error. We see that the simple strategy basically achieves this model limitation. 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 an in the example in table 1. If we assume the underlying dependence to have remained constant so that $\underline{E_0}$ 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 $\overline{\sigma^2}$). It appears from table 1 that of the four currencies, the British Pound's volatility seems to have increased while the remaining three markets display decreasing volatility, most notably that the German Mark.

5 Conclusion

We have shown how bounds on learning performance can be used in financial markets to obtain bounds on model limitation and to quantify the trade-off between numerous, more noisy data and fewer, less noisy data. Our results were applied to the currency markets to obtain estimates of the model limitations and to detect changes in volatility. They indicate that today's exchange rate comes close to being the best linear predictor of tomorrow's exchange rate.

References

1. F. Black and M. S. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 3:637–654, 1973.
2. M. Magdon-Ismail, A. Nicholson, and Y. S. Abu-Mostafa. Financial markets: very noisy information processing. *To appear In Proceedings of the IEEE Special Issue on Information Processing*, 1998.
3. N. Murata, S. Yoshizawa, and S. Amari. Learning curves, model selection and complexity of neural networks. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 607–614. Morgan Kaufmann, 1993.
4. J. Shao and D. Tu. *The Jackknife and the Bootstrap*. Springer-Verlag, New York, 1996.

This article was processed using the LaTeX macro package with LLNCS style