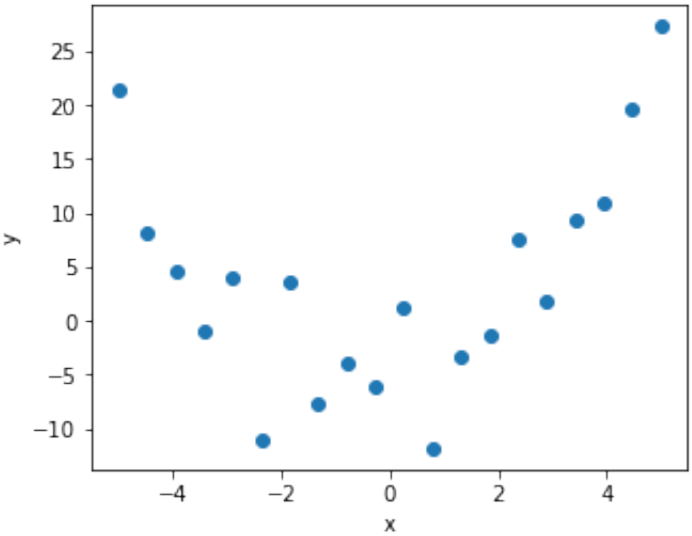
非線形な相関

線形相関で表せない場合は?

入力の変化が小さく、出力もまたそれに応じて少しだけ変化する場合には、線形フィッティングはうまくいきますが、入力の変化が大きくなるにつれて、グラフは曲がってきます。例えば、以下のようなデータが得られた時に、これを直線でフィットするのは勇気がいるでしょう。



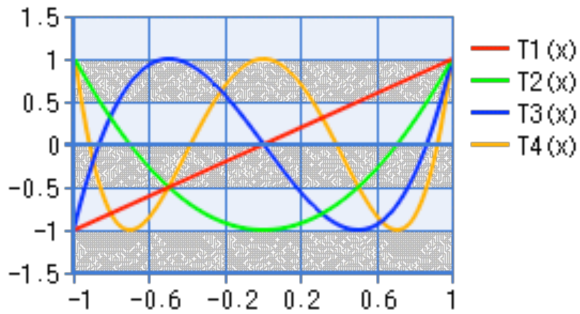
この場合には2次関数でフィットすれば良いかもしれません。ただし、一次であわないから二次にして、それでより近い線がひければめでたしめでたし、と考えるのは安易すぎます。三次でもなく一次でもなく二次が一番良いという、理論的な裏付けがある場合には躊躇なく二次関数を選べますが、裏付けがない場合には、逆にこのデータから、二次関数的な応答を生みだす理由(メカニズム、モデル)を見付けないと、線を引いただけでは「わかった」ことにはなりません。

多項式による漸進的フィッティング

それでも、何か線をひいてみることで、手がかりが得られるかもしれないので、次数を決めずにいろんな多項式でフィッティングしてみよう。

データを高次関数でいきなりフィットしようとすると、パラメータ空間が広すぎるために、よほど良い初期値を与えないと最適解を見付けられない可能性が高い。そこで、通常は低次関数から順に次数を上げていくという方法をとる。 しかし、1次から二次に次数を上げる際に、一次で得られたパラメータを、二次のパラメータの初期値としてどう使えばいいのかは、すぐにはわからない。 そこで便利なのが、Chevishev多項式 である。

$$T_0(x) = 1$$
$$T_1(x) = x$$
$$T_2(x) = 2x^2 - 1$$
$$T_3(x) = 4x^3 - 3x$$
$$T_4(x) = 8x^4 - 8x^2 + 1$$
$$T_5(x) = 16x^5 - 20x^3 + 5x$$
$$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1$$
$$T_7(x) = 64x^7 - 112x^5 + 56x^3 - 7x$$
$$T_8(x) = 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1$$
$$T_9(x) = 256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x$$
$$T_{10}(x) = 512x^{10} - 1280x^8 + 1120x^6 - 400x^4 + 50x^2 - 1$$
$$\vdots$$



1次〜4次のチェビシェフ多項式のグラフを図に示した。この多項式の特徴は、それぞれが互いに直交しているということである。

$$\int_{-1}^{1} T_m T_n(x) \mathrm{d}x = 0, m \neq n$$

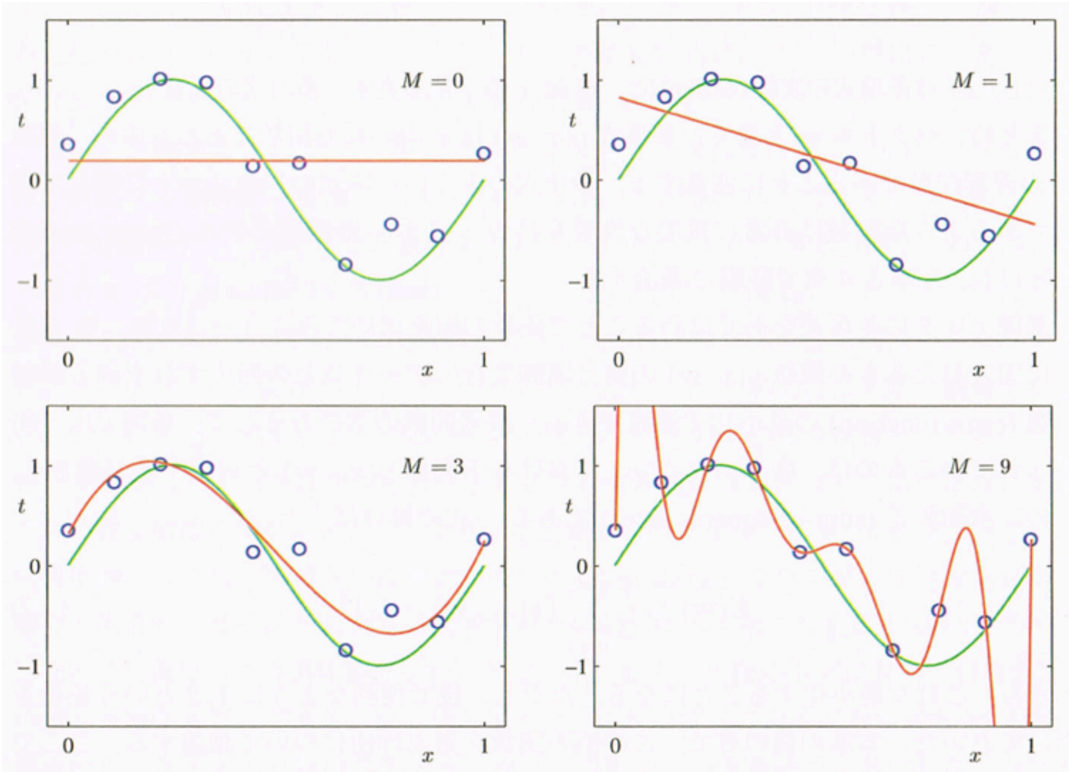
直交していると何がありがたいか? 区間 [−1, +1] で定義されるどんな関数 f(x) も、チェビシェフ多項式の線形和で一意的にあらわせるということだ。つまり、

$$f(x) = \sum_{i=0}^{\infty} a_i T_i(x)$$

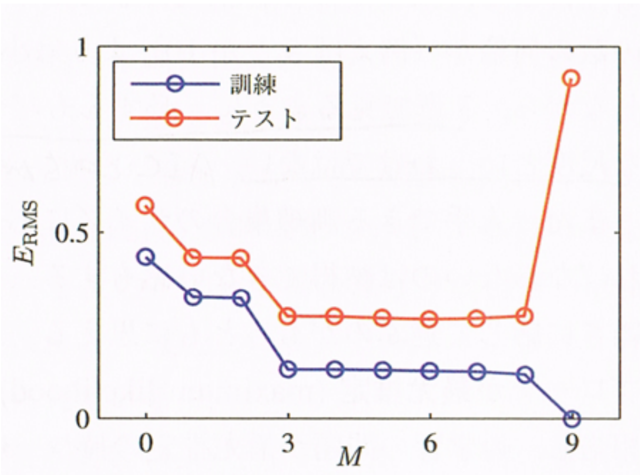
の形で一意的に展開できる、ということになる。 同じく互いに直交している、 $\sin nx$ と $\cos nx$ を使って、任意の関数 $f(x)$ がフーリエ級数展開できるのと同じ理屈である。

何次関数でフィットすれば十分と言えるか

サンプルデータとして、緑線(サイン曲線)にノイズがのったデータ(青丸)10個を用いる。 これを0次、1次、3次、9次関数でフィットした結果が下の図の赤線。 9次は「過学習」(overfitting)の状態。これを見ても、9次関数はやりすぎだと感じる。じゃあ、3次が良いのか、4次が良いのか、それとも5次か、というあたりはやはり微妙な判断になる。



そこで、元のデータ10個を訓練用データとし、それとは別にデータをもう10個準備して、これをテスト用データとする。下図は、フィッティング関数の次数に対し、訓練用データの誤差と、それを使ってテストデータを評価した誤差をプロットしている。9次関数は、訓練用データを完璧に通り、誤差は0だが、テストデータに対して破綻する、過学習の状態であることを示している。8次までの傾向をみると、3次関数よりも高次の関数を使っても、訓練データ、テストデータとも精度が上がらないことがわかる。このことから、上のグラフをフィットするには3次関数で十分であると言える。



ここでは次数を決定するのが目的なので、データセットをはじめにランダムに2つのセットに分け、半分をフィッティング用に、半分をテスト用に使って上のテストを行えばよい。次数が決まったら、すべてのデータをフィッティングに使用して、多項式の係数を決定する。(参考書: CMビショップ、「パターン認識と機械学習(上)」、丸善)

このように、訓練用データセットとテストデータセットを分ける方法は、ニューラルネットワークの学習過程でも広く利用されている。

```
In [ ]:
```