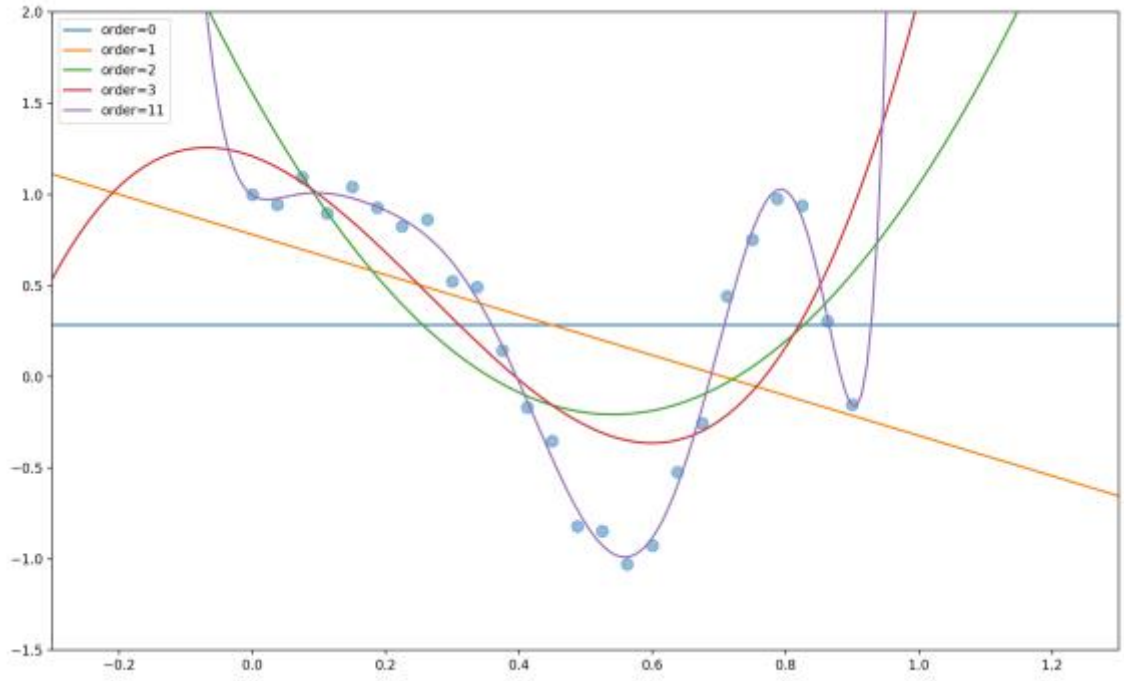
Coursework 1 Mathematics for Machine Learning (CO-496) CID:01366977

Name: Fanbo Meng

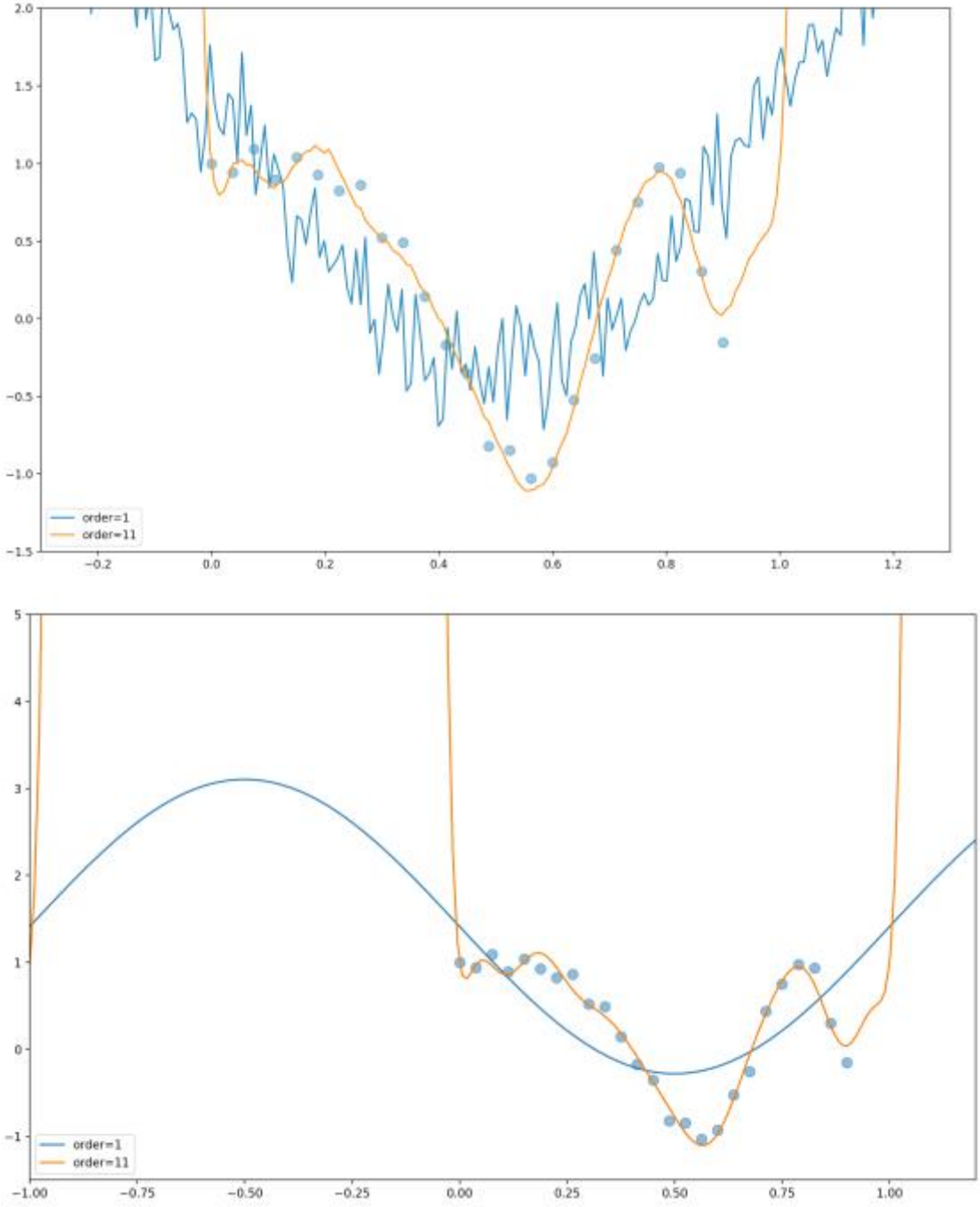
1. Linear Regression

a)



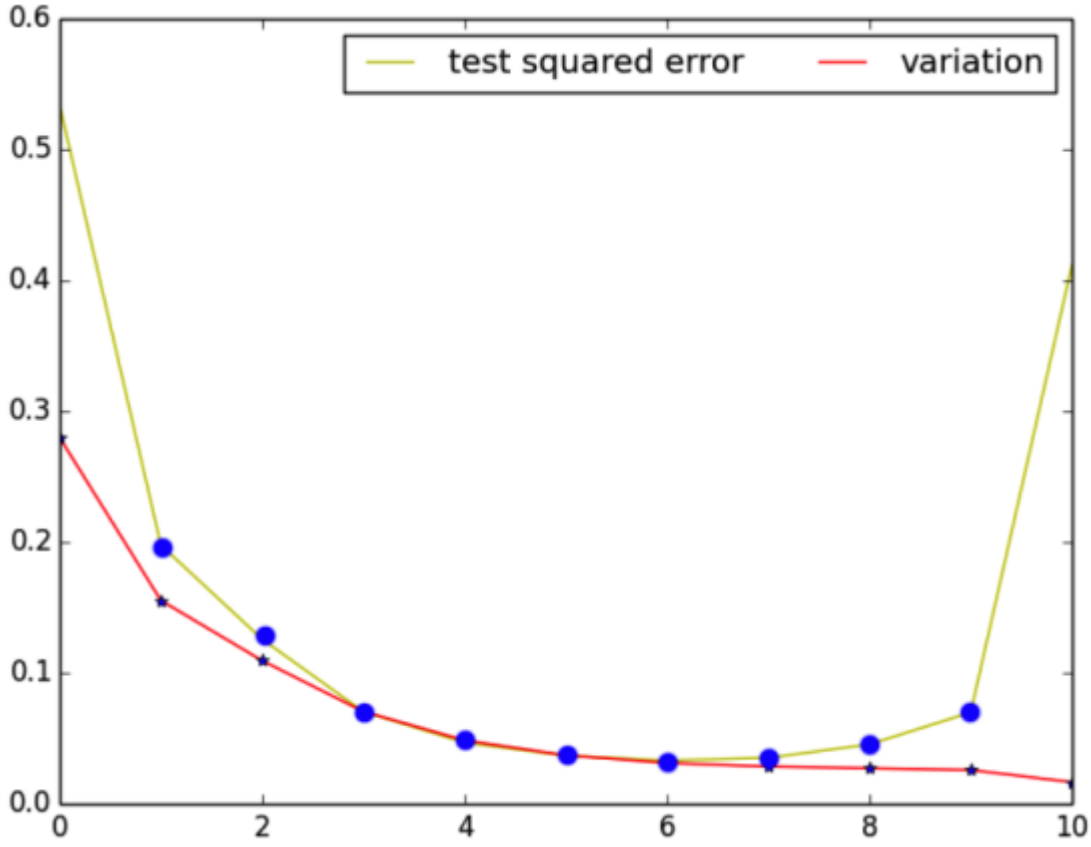
1.1Curves shown polynomial in order 0,1,2,3and 11

b)



1.2 Estimation with 1-order and 11-order trigonometric basis function

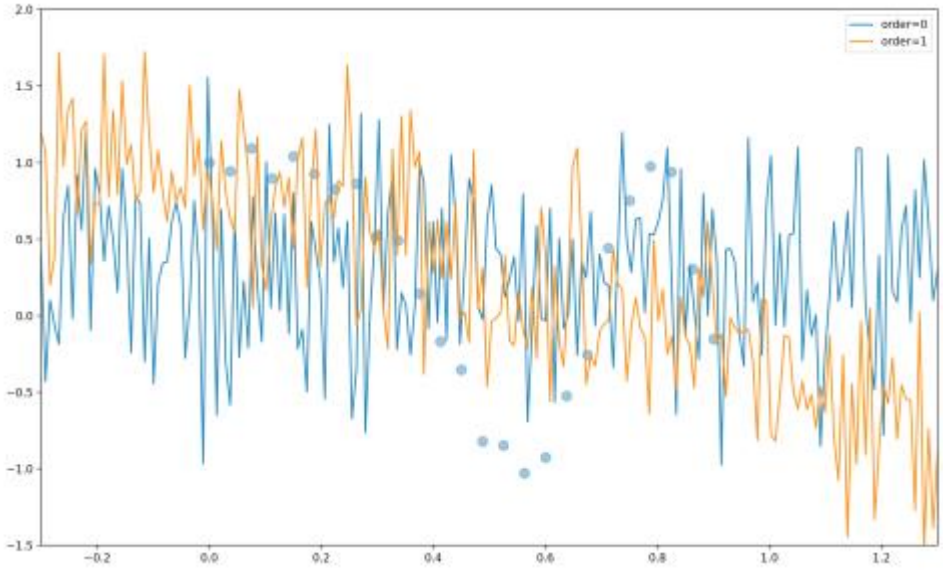
c)



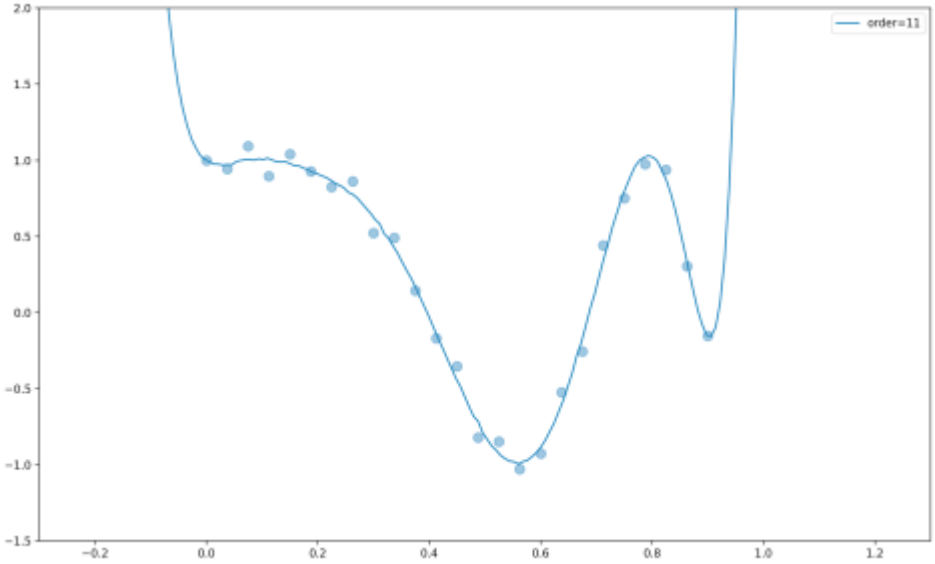
1.3 Curves for σ_{ML}^2 and test squared errors

d)

1. In Gaussian model, variances σ_{ML}^2 indicate how the data converges to means, in another word, every y satisfies a Gaussian distribution with a mean of $w^T \phi(x)$ and a variance σ^2, in which variance means that to what extent that y converges to $w^T \phi(x)$. When increasing the order of basis function and the weight vector w, since the basic function is getting more detailed with a higher order, estimated y will be more likely to approach $w^T \phi(x)$, so the variance σ_{ML}^2 is getting smaller. This can also be seen in graph 1.4 and 1.5, of which the curve of 11-order basis function goes significantly more smoothly than 1-order one.



1.4 Estimation with 0-order and 1-order polynomial basis functions



1.5 Estimation with 11-order polynomial basis function

2.In graph 1.3, we see that the square error with cross validation is generally getting smaller with the increasing order. This make senses, since with the higher order is, the more the model fits the data. However, we see that when the order increasing, the error raises dramatically. This is because with the order increasing, over-fitting problems start to show up.

3. Over-fitting is a problem that the model fits the training data well but has a poor job in prediction and testing data. This may happen when we training a model and estimating the parameters, the training data we use may contain

noises, so when the model try to fit all the data in training datasets, unnecessary noisy data is included. This may make the model so complicate that far from real.

2.Ridge Regression

a)

1.

Prior of w: w is an m-dimensional array, the elements of which (denoted by w_i) are i.i.d. Each of the elements is Gauss distribution. For convenience, I use this form to show the prior:

$$w_i \sim N(0, \sigma^2/\sqrt{\lambda})$$

w_i is a uniform notation for each component in vector w

2. $-logL_{MLE}(w,\sigma) =$

$-log\prod_{i=1}^{n}\left(\sqrt{2\pi}\sigma\right)^{-1}\exp\left(-\frac{(y_i-w^T\phi(x_i))^2}{2\sigma^2}\right)=\sum_{i=1}^{n} log\ (\sqrt{2\pi}\sigma) +$

$\sum_{i=1}^{n}\frac{(y_i-w^T\phi(x_i))^2}{2\sigma^2}$(this is the log likelihood)

$-logL_{MAP}(w,\sigma) = \sum_{i=1}^{n} log(\sqrt{2\pi}\sigma) + \sum_{i=1}^{n}\frac{(y_i-w^T\phi(x_i))^2}{2\sigma^2} + nlog(\sqrt{2\pi}\sigma/\sqrt{\lambda}) +$

$\lambda w^T w/2\sigma^2$ (In this problem I focus on posterior log function)

3.

Then we re-consider the loss function:

$$L(\boldsymbol{w}) = \sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\boldsymbol{\phi}(x_i))^2 + \lambda\sum_{j=1}^{M} w_j^2$$

We can see that given σ, the posterior function is the loss function produced by $2\delta^2$ and plus constants. So the differentials of both loss function and posterior function are only differs in parameters independent from w, which is:

$$\frac{dL}{dw_j} \propto \frac{dL_{MAP}}{dw_j} \propto \sum\nolimits_{i=1}^{N} -\phi_j(x_i)(y_i - w^T\phi(x_i)) + \lambda w_j$$

4. /5.

The intuition of loss function is that we can calculate the w with prior (the previous approximation before we get the training data) and the likelihood. As I

indicated in the first part, λ is a parameter indicating how much contribution the prior make to the final estimation. Since prior is Gaussian distribution, with a mean of $\mathbf{0}$, and variance of $\sigma^2/\sqrt{\lambda}$, in this way, λ controls the variance of prior, when λ is big, it means that the variance of prior is small, shows that previous knowledge about *w* is confident, so more contribution is made by prior, and if the λ is small, that the previous knowledge about *w* is less confident, so little contribution is made by prior. This logic is similar with that of MAP estimation, since according to Bayes, posterior is a combination of prior and likelihood, that is why we get the same solution for *w* by using loss function and MAP estimation.

b) Here is the graph showing different λ causing different results.