LECTURE 18 - PREDICTIVE DISTRIBUTION

- Ok, so in all of this regression discussion we have had, we've focused on estimating $\mathbf{w}$. But, we really dont care about the value of $\mathbf{w}$, actually.
- What we care about is predicting the correct value of t for new incoming test data.
- One way to do that, as we have discussed so far, is estimating a single $\mathbf{w}$ from training data and then using that to estimate t on incoming test data.
- But we can go beyond this to also estimate the *predictive distribution* on t:

$$p(t|\mathbf{t},\alpha,\beta) = \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)d\mathbf{w} \quad (1)$$

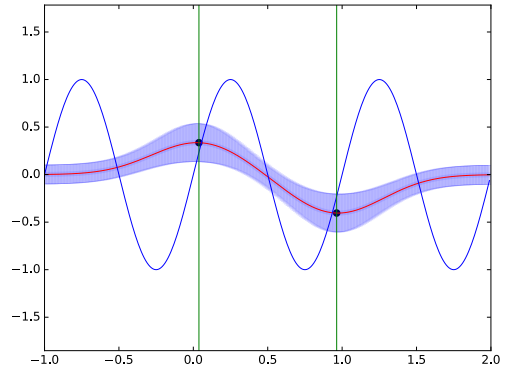
$$= \int \mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)d\mathbf{w} \quad (2)$$

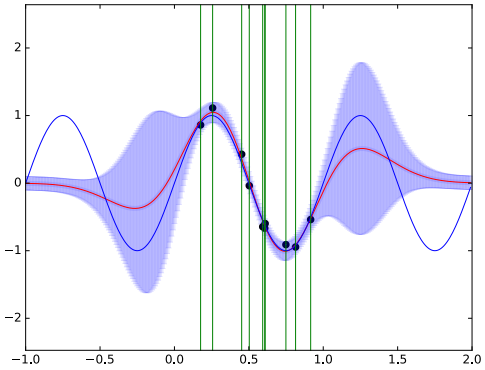
$$p(t|\mathbf{t},\alpha,\beta) = \mathcal{N}(t|\mathbf{m}_n^T\mathbf{\Phi}(\mathbf{x}),\frac{1}{\beta}+\mathbf{\Phi}(\mathbf{x})^T\mathbf{S}_N\mathbf{\Phi}(\mathbf{x})) \quad (3)$$

- *How do we get (3) from (2)*? This is the marginal distribution of t. We will take it as a given for now.
- *How would you interpret Equation 1? What does the predictive distribution mean?*
- *How would you interpret the two terms of the variance in the predictive distribution?* noise on data, uncertainty associated with $\mathbf{w}$
- As additional data points are observed, the variance for the predictive distribution becomes narrower.
- In all of the following discussion, we assumed Radial Basis kernels centered at the data.
- *What happens to the predictive variance at a data point?*

  As can be seen in the figure below, the predictive variance at a data point tends to decrease. This is desireable because since there is data at this point, we have knowledge about the desired values in this region of the feature space.
- *What happens to the predictive variance as you move away from a data point?* As can be seen in the figure, the predictive variance increase as you move slightly away from regions of data. This is desireable because since there is no data at this point, we do not have a lot of knowledge about the desired values in this region of the feature space.
- *What happens to the predictive variance far, far from any data?* As can be seen in the figure, the predictive variance decreases and falls to the prior precision value as you move far away from regions of data. This is NOT desireable because since there is no data at this point, we do not have a lot of knowledge about the desired values in this region of the feature space yet the variance indicates certainty in the solution through lack of variance.

- *How well do you approximate the true function as you increase the number of data points?*





As can be seen by comparing the two figures, the true function is better approximated with more data.

Run the provided code and discuss the following questions:

(1) *Run the provided code and recreate the figure below. What is this showing? How do parameter settings change the results?*

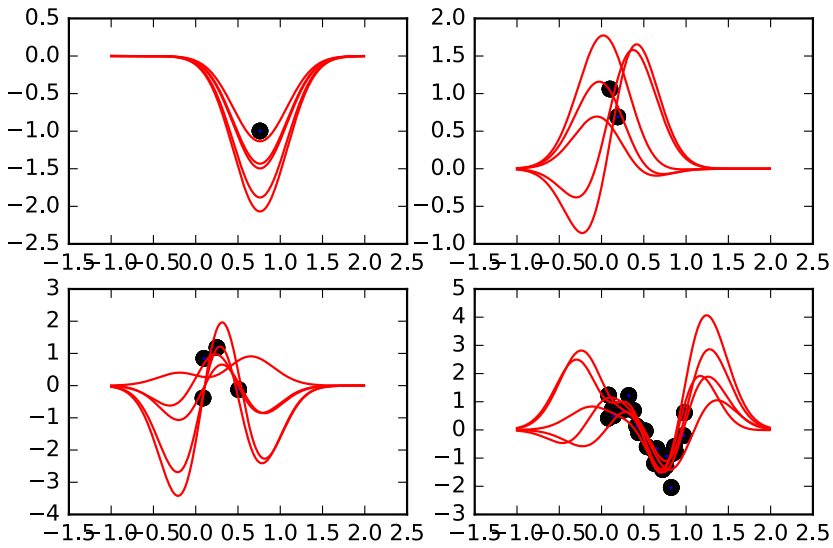
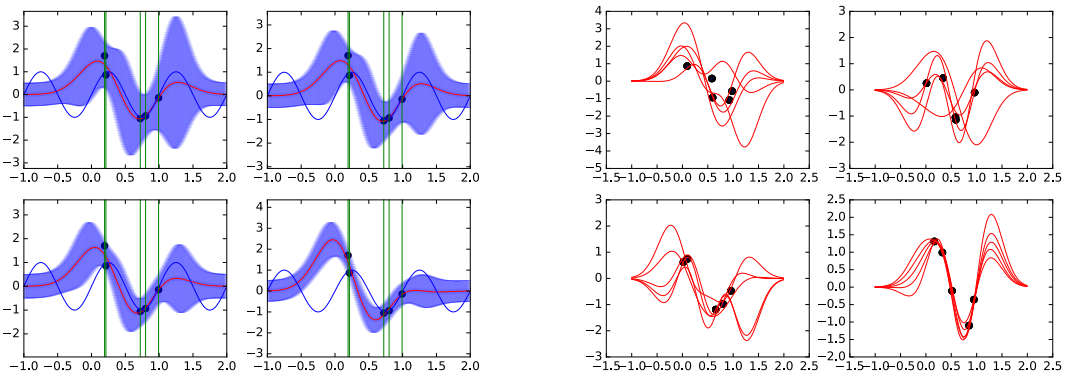


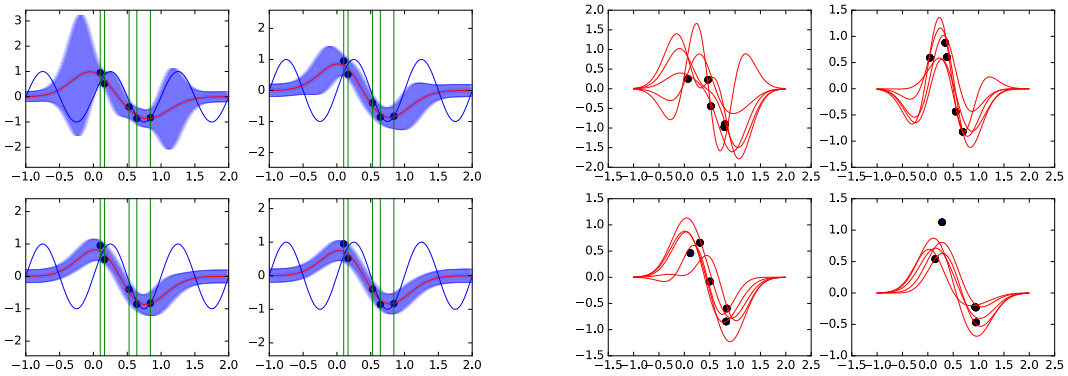
Figure above shows plots of $y = \mathbf{w}^T\mathbf{x}$ over the range [-2,2] using five samples of $\mathbf{w}$ drawn from the posterior distribution over $\mathbf{w}$ given one, two, four and 25 data points. As can be seen in the figure, the five samples agree more when there are more data points (i.e., the variance of the posterior distribution decreases) and provides a better fit of the data. Also, as can be seen in the figures, the variance of the plots decrease as you move far from the data (relative to the width of the RBF basis functions) which may be undesireable (indicating certainty in the function there when there should be none).

(2) *Vary the parameter value. Discuss the role that plays and how it effects the predictive distrubtion.*



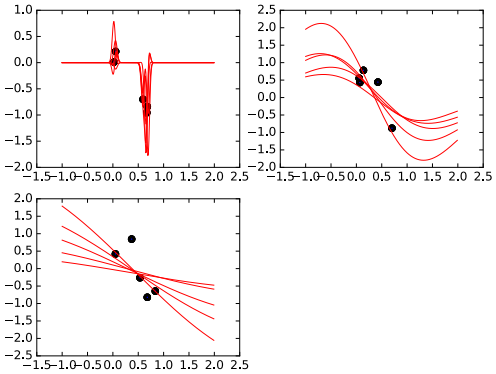
The two figures above show the predictive distribution and plots given samples from the predictive distribution given β values of 1,2,10, and 100, respectively. As can be seen in the plots, β controls the predictive variance. Large β values result in a smaller predictive variance.

(3) *Vary the α parameter value. Discuss the role that plays and how it effects the predictive distrubtion.*



The two figures above show the predictive distribution and plots given samples from the predictive distribution given α values of 0.001,0.01,0.1, and 1, respectively. As can be seen in the plots, α is the prior precision and effects the importance of the prior belief in the prediction and also plays a role in the resulting predictive distribution variance.

(4) *Vary the s parameter value. Discuss the role that s plays and how it effects the predictive distrubtion.*



The figure above show plots given samples from the predictive distribution given s values of 0.001,1, and 10, respectively. As can be seen in the plots, s controls the width of the RBF basis functions. Too small of an s value and

prediction can only be done very locally, too broad and there is poor fit of the data since each RBF influences too large of a range of values.