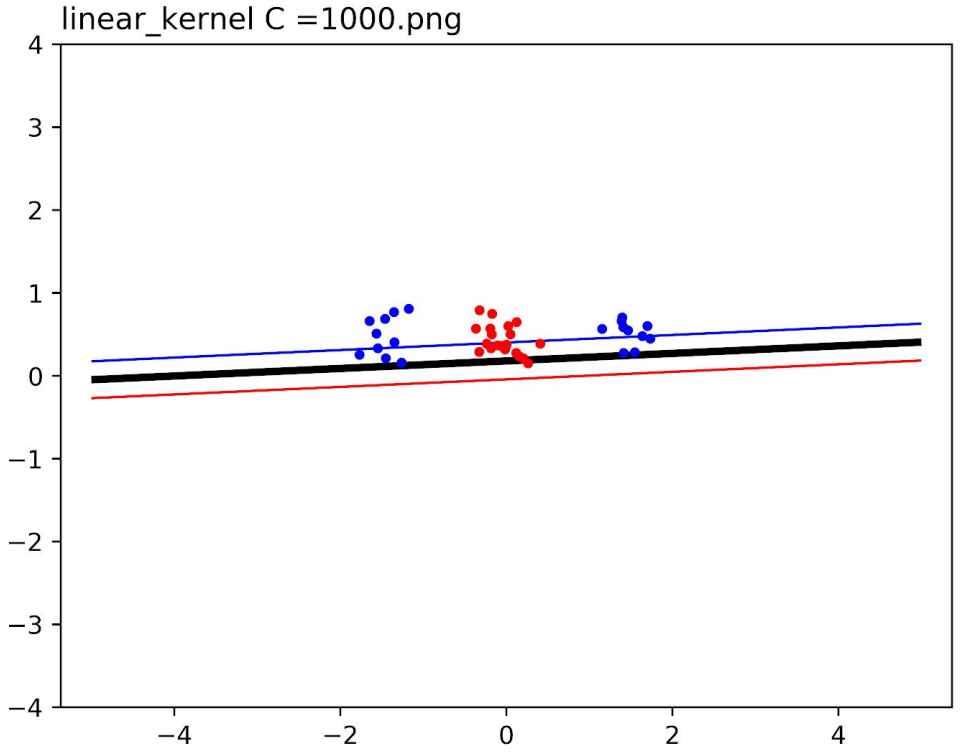
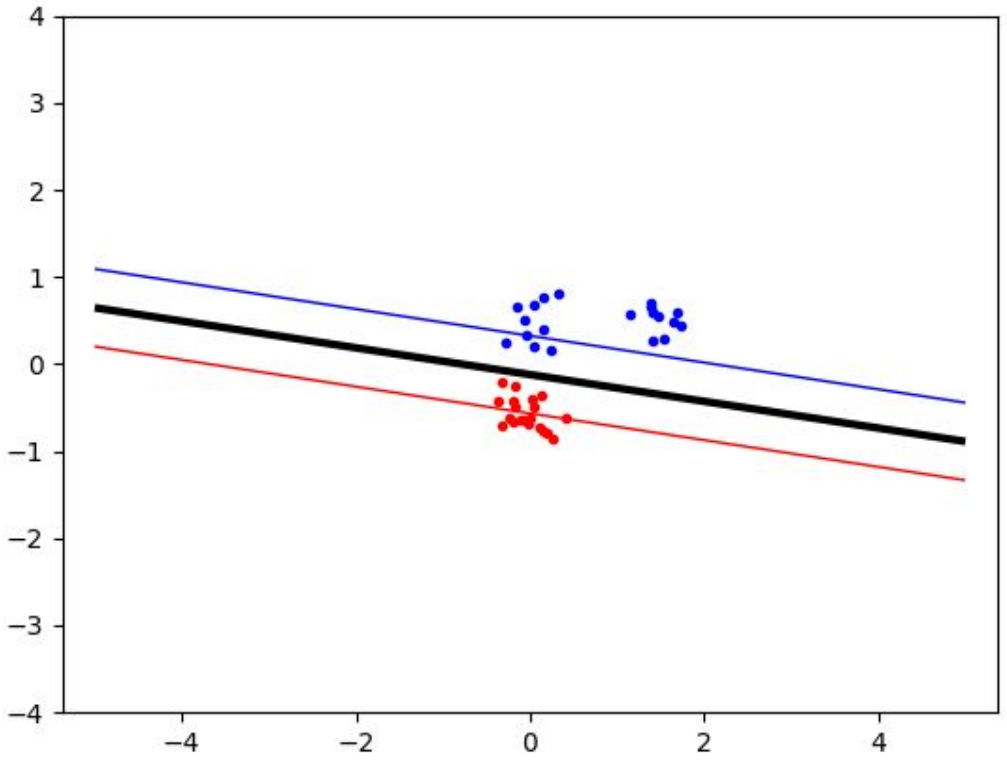
Nonseparability and separability



The dataset above is not linearly separable, optimizer returns False.



The above dataset is linearly separable.

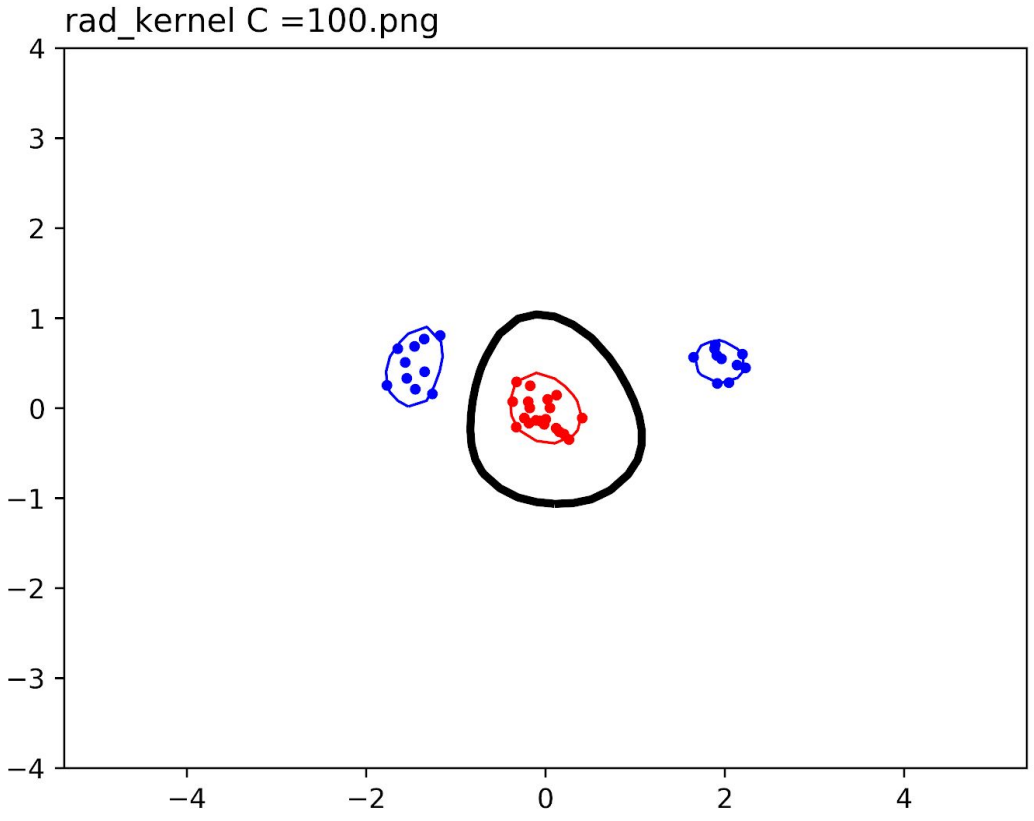
Sigma

Parameters

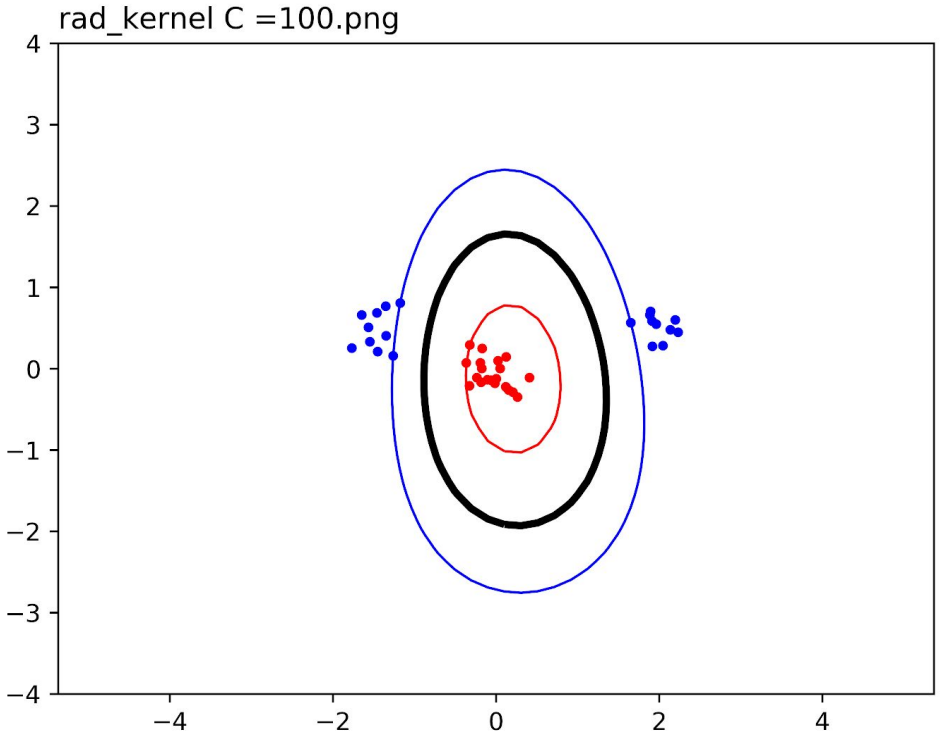
With a small sigma, the boundary and margins is well fitted to the data.

With a large sigma, the boundary and margins is less fitted to the data.

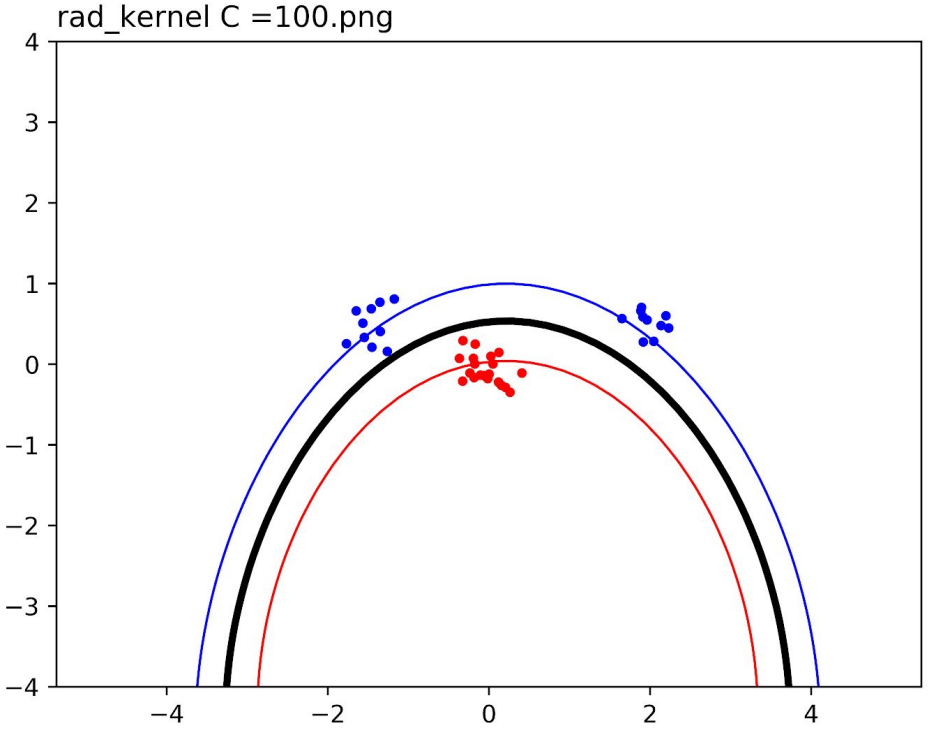
In terms of bias-variance trade off, a smaller sigma results in higher variance, since it is fitted to the training data well. The tradeoff is that the bias become smaller.



Sigma = 0.5



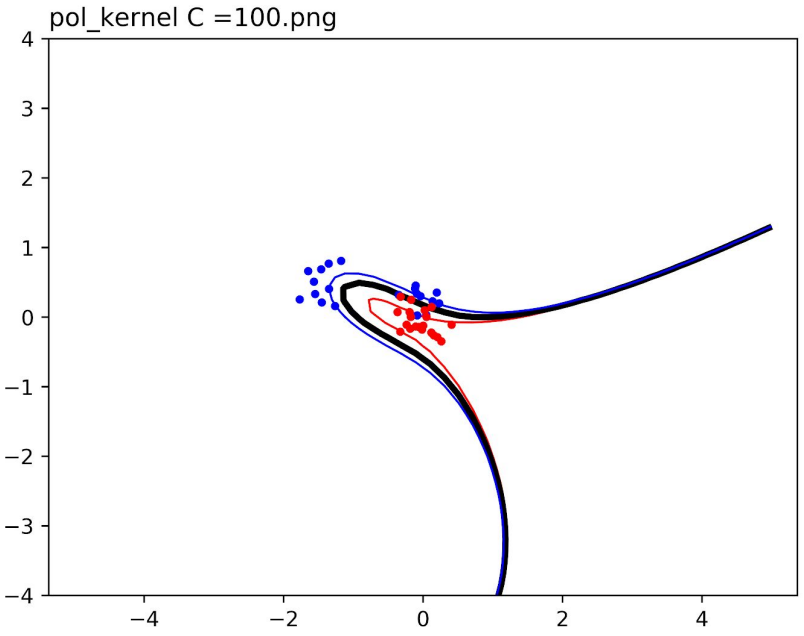
Sigma =2



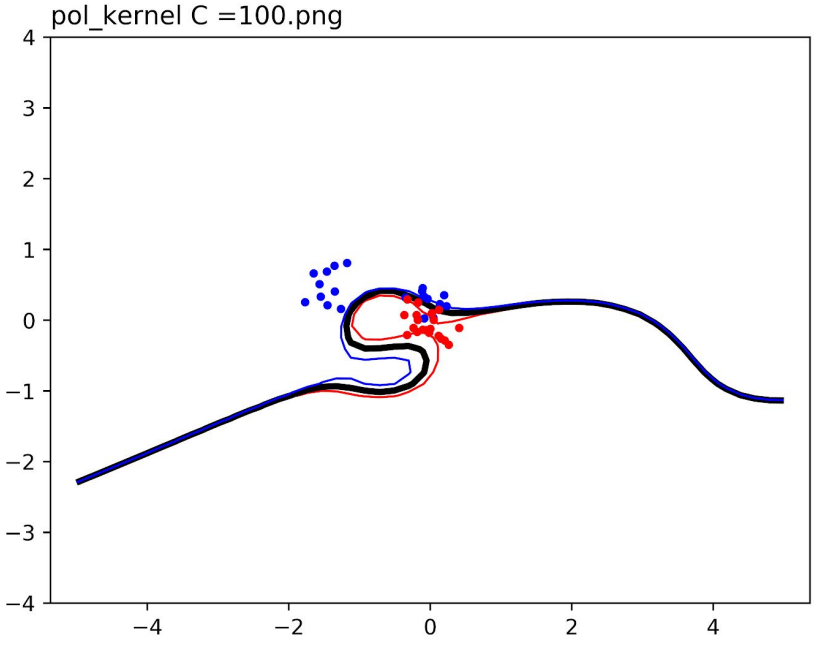
Sigma = 7

Polynomial p

With a higher polynomial p the result is more complex shapes. This means that the shape gets a tighter fit to the training data, thus a higher p results in higher variance. However, the tradeoff is that the bias becomes lower.



p = 3

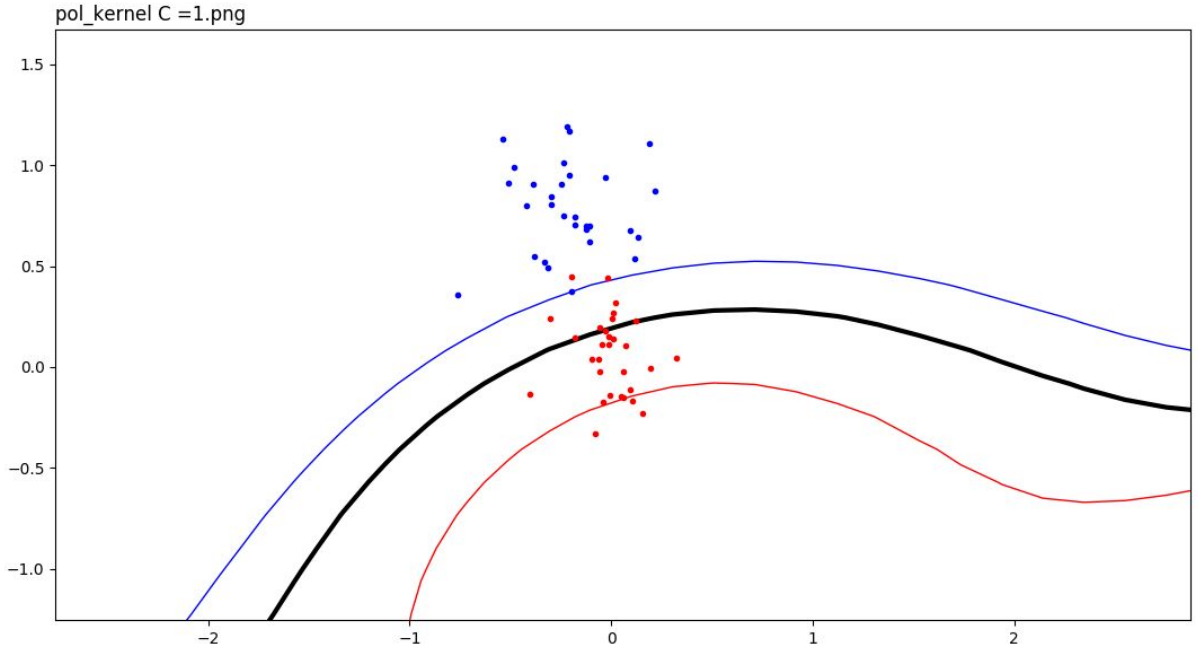


p= 5

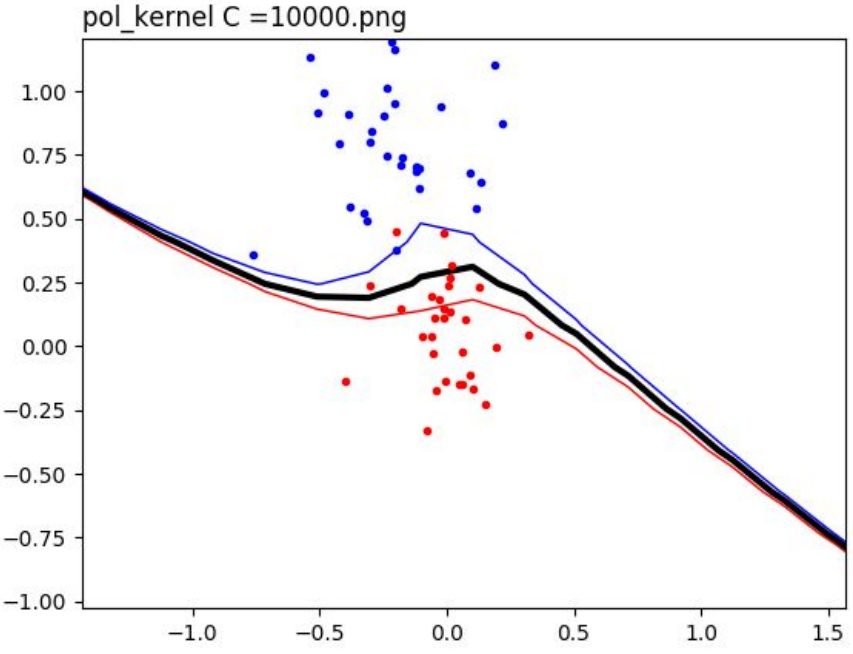
Slack parameter C

A lower C value means that the “budget” for misclassification is higher. That is, we can have (more) points on the wrong side of the boundary and inside the margins as well.

A high C value means that we have a very low budget for misclassification.



With C = 1, we get 7 points on the wrong side of the boundary.



With C= 10000, we get 4 points on the wrong part of the boundary as well as a few points inside the margin. This makes sense, since points on the right side of the boundary but on the wrong side on the margin should be less costly than points on the wrong side of the boundary.

In terms of variance bias tradeoff, a lower C value should result in less variance since the boundary is not fitted as tight to the data but instead allow for some, well, slack. The tradeoff is that the bias should be higher.

Imagine

The goal with statistical learning as we understand it, and according to the book p. 17, is to try to estimate a function f^ that estimates the true underlying function f.

If we suspect that the underlying function is linear in nature, but not so easily separable, it would make sense to increase slack but keep a less complex model. However, if we think that the underlying function is not linear it makes more sense to have a more complex kernel with less slack.