<u>The Bias-Variance trade-off</u>

I explain why the bias-variance trade-off happens based on the expected test MSE.

The expected test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0, is given by

$$\mathrm{E}\left[\left(\hat{f}\left(x_0\right)-y_0\right)^2\right]=\mathrm{Var}\left(\hat{f}\left(x_0\right)\right)+\left(\mathrm{E}\left(\hat{f}\left(x_0\right)\right)-y_0\right)^2+\mathrm{Var}\left(\epsilon\right), \tag{1}$$

where the second term of the L.H.S. of the Eq.(1) is called *bias* of the estimator $\hat{f}$.
As we said, the expectation here is taken in relation to the possible training sets, i.e. the training sets available are taken as a sample. Therefore the expectation is not taken in relation to X as it seems to be by the notation.
We often hear hear that there is a trade-off between bias and variance of $\hat{f}$, but where does it come from? At a first glance, it seems that a greater variance implies a greater bias as the error $\left(\hat{f}\left(x_0\right)-y_0\right)$ can be greater. But that is the point, the error *can* be greater but can also be smaller. And for a highly non-linear response y an inflexible method leads, in general, to greater errors $\left(\hat{f}\left(x_0\right)-y_0\right)$ than an flexible method leads. Of course, if the response is linear, a inflexible method such as a linear regression will have a smaller bias and, as in general, a smaller variance than flexible methods. But this is an exception, and in general we have this trade-off.

Why are we more interested in the expected test MSE than in bias?

The expected test MSE help us better in the decision of what statistical method should we choose in order to predict/describe a certain problem because the bias does not take into account the effects of the variance of $\hat{f}$ but only its average at x_0. If a method estimates some $\hat{f}$ whose error $\left(\hat{f}\left(x_0\right)-y_0\right)$ is huge, this will be taken into account by $\mathrm{E}\left[\left(\hat{f}\left(x_0\right)-y_0\right)^2\right]$, by its own definition, but it will not be taken into account by the bias $\left(\mathrm{E}\left(\hat{f}\left(x_0\right)\right)-y_0\right)^2$ as only the average value $\mathrm{E}\left(\hat{f}\left(x_0\right)\right)$ matters.