After only 20 examples, your learned function will not yet look like the target function. Explain in a paragraph why it looks the way it does. If your learned function involves many peaks and valleys, then be sure to explain both their number, their height, and their width.

When learning starts, the 3D function graph would look flat since the estimates for weights are all 0. From here, I like to consider the effect of new, incoming training data. Processing training data (a new input/output example) is visualized by 1. pinching the graph at the input value and then 2. pulling it up (or down) based on the output value. Depending on what the step size is, the graph will stretch part or all of the way to where you stretch it. ie. a step size of 1 will cause the graph to "stick" where you let go. While 0.5 will result in it recessing 0.5 of the way back. Also, it isn't just the point at which you pinch the graph that is stretched. It's the area around it. The "area around it" is a function of what the symmetry of the tilings/features. Wide tilings mean that the tiles are wide, and thus, a wide range of surrounding areas of the graph are pulled up with the pinch. With this model in mind, it is easy to see that each bit of training data will cause a "peak" or value. The height depends on what the target function says the value should be (and how often it's been pulled in that area of the graph). Obviously if input values are close together, they generalize and the peaks combine so that it may look like one peak (or valley). Also, if the target function returns 0, the training data wouldn't change the graph much, so neither peak or valley would be obvious.

Suppose that instead of the tiling input space into an 11X11 grid of squares, you had divided into an 11X21 grid of rectangles with the in1 dimension being divided twice as finely as the in2 dimension. Explain how you would expect the function learned after 20 examples to change if this alternative tiling were used.

An 11X11 grid is uniformly offset tilings. For these types of offsets, generalizations happen along the diagonal of input space. With asymmetric offsets (which would be the case with an 11X21 grid), the generalizations are more spherical. Furthermore, if the dimension is defined more finely, the generalizations won't be so broad. Therefore, the peaks and valleys on the graph won't be as wide. Furthermore, where two different inputs in the past may have combined to form a peak, they may now form 2 distinct peaks since they no longer overlap.