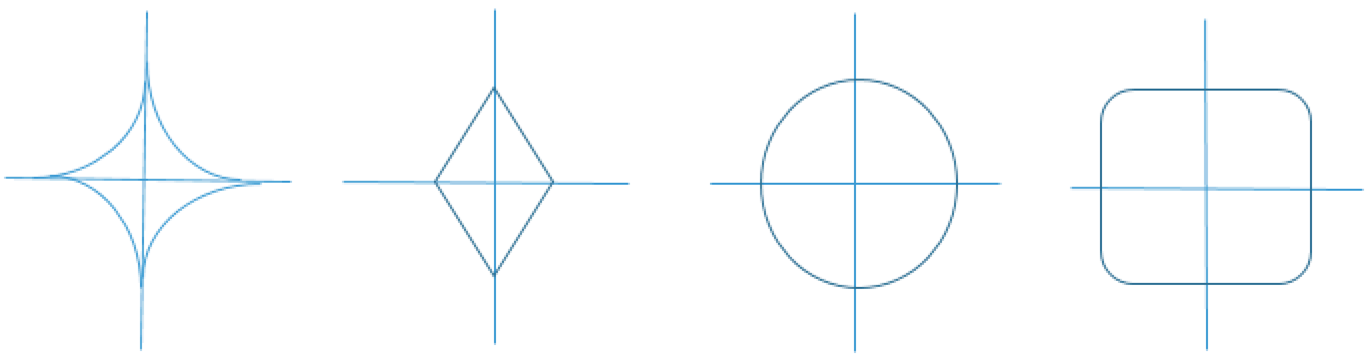
# L1 & L2 Regularization

Actually, there are different possible choices of regularization with different choices of order of the parameter in the regularization term, which is denoted by $\sum_i |\theta_i|^p$ . This is more generally known as $L_p$ regularizer.

Let us try to visualize some by plotting them. For making visualization easy, let us plot them in 2D space. For that we suppose that we just have two parameters. Now, let's say if p=1, we have term as $\sum_i |\theta_i|^p = |\theta_1| + |\theta_2|$.. Can't we plot this equation of line? Similarly plot for different values of p are given below.



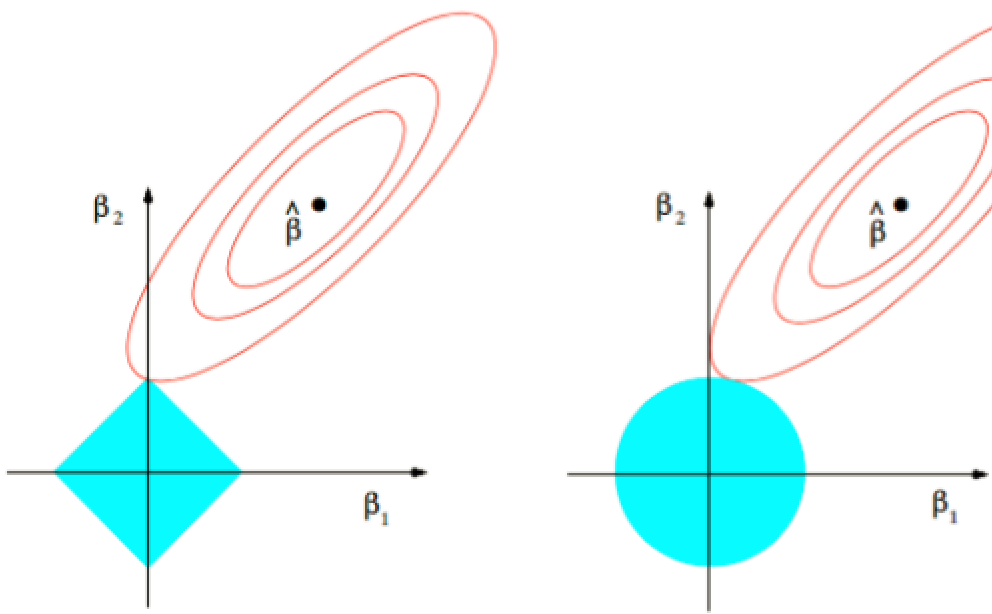
p = 0.5 p =1 p = 2 p = 4

In the above plots, axis denote the parameters($\Theta_1$ and $\Theta_2$). Let us examine them one by one.

For p=0.5, we can only get large values of one parameter only if other parameter is too small. For p=1, we get sum of absolute values where the increase in one parameter Θ is exactly offset by the decrease in other. For p =2, we get a circle and for larger p values, it approaches a round square shape.

The two most commonly used regularization are in which we have p=1 and p=2, more commonly known as L1 and L2 regularization.

Look at the figure given below carefully. The blue shape refers the regularization term and other shape present refers to our least square error (or data term).



The first figure is for L1 and the second one is for L2 regularization. The black point denotes that the least square error is minimized at that point and as we can see that it increases quadratically as we move from it and the regularization term is minimized at the origin where all the parameters are zero .

Now the question is that at what point will our cost function be minimum? The answer will be, since they are quadratically increasing, the sum of both the terms will be minimized at the point where they first intersect.

Take a look at the L2 regularization curve. Since the shape formed by L2 regularizer is a circle, it increases quadratically as we move away from it. The L2 optimum(which is basically the intersection point) can fall on the axis lines only when the minimum MSE (mean square error or the black point in the figure) is also exactly on the axis. But in case of L1, the L1 optimum can be on the axis line because its contour is sharp and therefore there are high chances of interaction point to fall on axis. Therefore it is possible to intersect on the axis line, even when minimum MSE is not on the axis. If the intersection point falls on the axes it is known as sparse.

Therefore L1 offers some level of sparsity which makes our model more efficient to store and compute and it can also help in checking importance of feature, since the features that are not important can be exactly set to zero.