Train loss vs Iteration for different num_layers (dim=2)