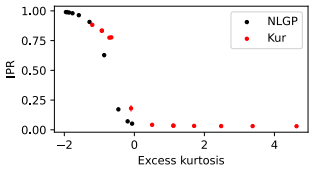

Figure 1: Inverse participation ratio (IPR) vs. excess kurtosis for `NLGP` and `Kur` datasets, plotted as mean $\pm$ std. dev. across 30 re-initializations, for the single-neuron model; error bars are small and may not be visible. IPR decays from its maximum of 1.0, indicating localization, to near zero as excess kurtosis increases. Effects on IPR across `NLGP` and `Kur` demonstrate independence of distributional properties.

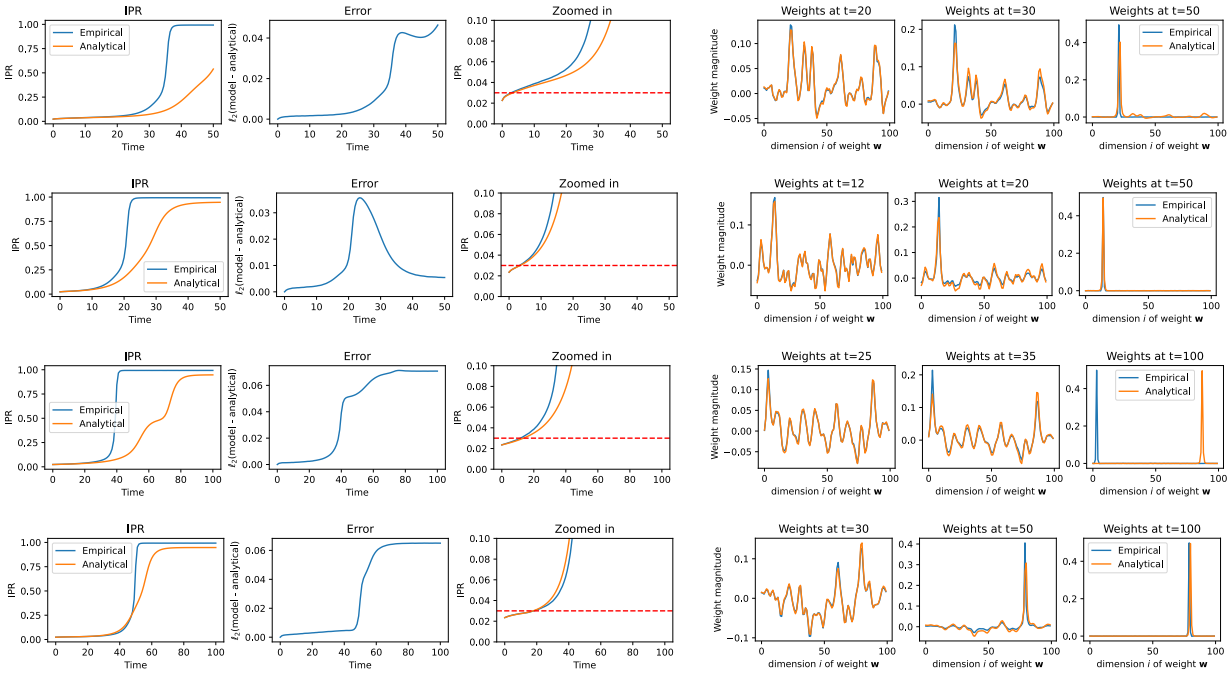

Figure 2: Four initializations trained on $\texttt{NLGP}(g = 100)$ with $\xi_0 = 0.3$ and $\xi_1 = 0.7$. As expected, weights always localize. In (Left, First) we plot IPR for empirical and analytical receptive fields (RFs) across time (defined as (# of gradient steps) $\times$ τ, the learning rate). In (Left, Second) we plot the time-evolution of ℓ_2 distance between the empirical and analytical RFs. In (Left, Third) we zoom in on (Left, First), restricting the range to $[0, 0.1]$ to more closely see divergence in IPR early in training. In (Right, First) and (Right, Second), we snapshot the empirical and analytical RFs at a time *before* and *just after*, respectively, the analytical model breaks down (according to IPR and ℓ_2 distance) due to localization. Finally, in (Right, Third), we snapshot *at the end* of the training period. In all but the third row, the analytical predictions are near-exact; in the third row, we predict localization, but at the wrong position. Focusing again on the first row, we see that at $t = 20$, the weights have not yet become localized (from IPR in (Left, First) and visually) and analytical and empirical weights match near exactly (confirmed by small distance in (Left, Center) above). At $t = 30$, a localized bump around $i = 21$ begins to emerge, violating assumption (A3) and weakening analytical precision. The analytical model then underestimates the degree to which the main bump at $i = 21$ dominates, while it overestimates the size of competing bumps at $i = 30$, 37, and 90. Despite this, at $t = 50$, we see that predictions from the analytical model retain a match to the empirical model.

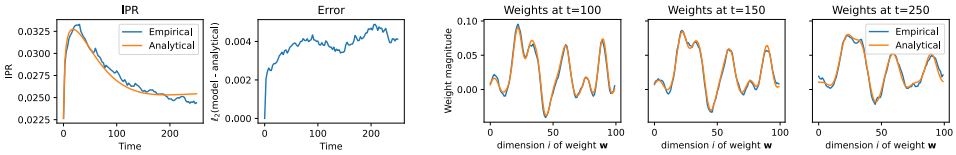

Figure 3: Same initialization as first row, but trained on $\texttt{NLGP}(g = 0.01)$ data, again with $\xi_0 = 0.3$ and $\xi_1 = 0.7$. As expected, weights do not localize. We plot the same quantities as above, but here the predictions of our analytical model hold *throughout* the entire training process as localization never emerges and so assumption (A3) is not violated as above.

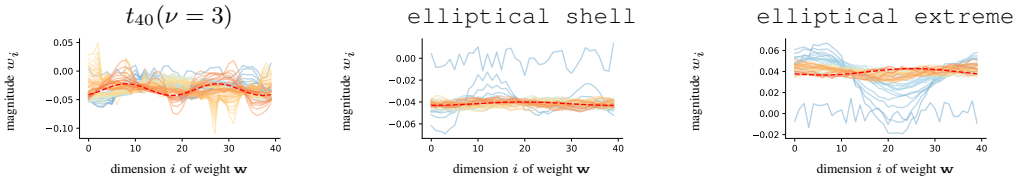

Figure 4: Receptive fields (RFs) trained on elliptical distributions, with MSE-fitted sinusoids (red dashes).This is to clarify our claim in proposition 3.3. The ℓ_2 distances between the fitted oscillatory weights and empirical RFs, as a ratio of the ℓ_2 norm of the empirical RFs, are (left) 9.77%, (center) 3.75%, and (right) 4.14%. (Left) appears different from the corresponding version in the manuscript; we used the same initialization, but scaled it down to be comparable to (middle) and (right) to reveal sinusoidal structure of the final RF.