



Figure 10.15: When composing many nonlinear functions (like the linear-tanh layer shown here), the result is highly nonlinear, typically with most of the values associated with a tiny derivative, some values with a large derivative, and many alternations between increasing and decreasing. In this plot, we plot a linear projection of a 100-dimensional hidden state down to a single dimension, plotted on the y -axis. The x -axis is the coordinate of the initial state along a random direction in the 100-dimensional space. We can thus view this plot as a linear cross-section of a high-dimensional function. The plots show the function after each time step, or equivalently, after each number of times the transition function has been composed.

1994; Pascanu *et al.*, 2013). In this section, we describe the problem in more detail. The remaining sections describe approaches to overcoming the problem.

Recurrent networks involve the composition of the same function multiple times, once per time step. These compositions can result in extremely nonlinear behavior, as illustrated in figure 10.15.

In particular, the function composition employed by recurrent neural networks somewhat resembles matrix multiplication. We can think of the recurrence relation

$$\mathbf{h}^{(t)} = \mathbf{W}^\top \mathbf{h}^{(t-1)} \quad (10.36)$$

as a very simple recurrent neural network lacking a nonlinear activation function, and lacking inputs \mathbf{x} . As described in section 8.2.5, this recurrence relation essentially describes the power method. It may be simplified to

$$\mathbf{h}^{(t)} = (\mathbf{W}^t)^\top \mathbf{h}^{(0)}, \quad (10.37)$$

and if \mathbf{W} admits an eigendecomposition of the form

$$\mathbf{W} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top, \quad (10.38)$$