

expresses our belief that the task should be accomplished via a multi-step program, with each step referring back to the output of the processing accomplished via previous steps.

- *Shared factors across tasks:* In the context where we have many tasks, corresponding to different y_i variables sharing the same input \mathbf{x} or where each task is associated with a subset or a function $f^{(i)}(\mathbf{x})$ of a global input \mathbf{x} , the assumption is that each y_i is associated with a different subset from a common pool of relevant factors \mathbf{h} . Because these subsets overlap, learning all the $P(y_i | \mathbf{x})$ via a shared intermediate representation $P(\mathbf{h} | \mathbf{x})$ allows sharing of statistical strength between the tasks.
- *Manifolds:* Probability mass concentrates, and the regions in which it concentrates are locally connected and occupy a tiny volume. In the continuous case, these regions can be approximated by low-dimensional manifolds with a much smaller dimensionality than the original space where the data lives. Many machine learning algorithms behave sensibly only on this manifold (Goodfellow *et al.*, 2014b). Some machine learning algorithms, especially autoencoders, attempt to explicitly learn the structure of the manifold.
- *Natural clustering:* Many machine learning algorithms assume that each connected manifold in the input space may be assigned to a single class. The data may lie on many disconnected manifolds, but the class remains constant within each one of these. This assumption motivates a variety of learning algorithms, including tangent propagation, double backprop, the manifold tangent classifier and adversarial training.
- *Temporal and spatial coherence:* Slow feature analysis and related algorithms make the assumption that the most important explanatory factors change slowly over time, or at least that it is easier to predict the true underlying explanatory factors than to predict raw observations such as pixel values. See section 13.3 for further description of this approach.
- *Sparsity:* Most features should presumably not be relevant to describing most inputs—there is no need to use a feature that detects elephant trunks when representing an image of a cat. It is therefore reasonable to impose a prior that any feature that can be interpreted as “present” or “absent” should be absent most of the time.
- *Simplicity of Factor Dependencies:* In good high-level representations, the factors are related to each other through simple dependencies. The simplest