

### 17.5.1 Tempering to Mix between Modes

When a distribution has sharp peaks of high probability surrounded by regions of low probability, it is difficult to mix between the different modes of the distribution. Several techniques for faster mixing are based on constructing alternative versions of the target distribution in which the peaks are not as high and the surrounding valleys are not as low. Energy-based models provide a particularly simple way to do so. So far, we have described an energy-based model as defining a probability distribution

$$p(\mathbf{x}) \propto \exp(-E(\mathbf{x})). \quad (17.25)$$

Energy-based models may be augmented with an extra parameter  $\beta$  controlling how sharply peaked the distribution is:

$$p_\beta(\mathbf{x}) \propto \exp(-\beta E(\mathbf{x})). \quad (17.26)$$

The  $\beta$  parameter is often described as being the reciprocal of the **temperature**, reflecting the origin of energy-based models in statistical physics. When the temperature falls to zero and  $\beta$  rises to infinity, the energy-based model becomes deterministic. When the temperature rises to infinity and  $\beta$  falls to zero, the distribution (for discrete  $\mathbf{x}$ ) becomes uniform.

Typically, a model is trained to be evaluated at  $\beta = 1$ . However, we can make use of other temperatures, particularly those where  $\beta < 1$ . **Tempering** is a general strategy of mixing between modes of  $p_1$  rapidly by drawing samples with  $\beta < 1$ .

Markov chains based on **tempered transitions** (Neal, 1994) temporarily sample from higher-temperature distributions in order to mix to different modes, then resume sampling from the unit temperature distribution. These techniques have been applied to models such as RBMs (Salakhutdinov, 2010). Another approach is to use **parallel tempering** (Iba, 2001), in which the Markov chain simulates many different states in parallel, at different temperatures. The highest temperature states mix slowly, while the lowest temperature states, at temperature 1, provide accurate samples from the model. The transition operator includes stochastically swapping states between two different temperature levels, so that a sufficiently high-probability sample from a high-temperature slot can jump into a lower temperature slot. This approach has also been applied to RBMs (Desjardins *et al.*, 2010; Cho *et al.*, 2010). Although tempering is a promising approach, at this point it has not allowed researchers to make a strong advance in solving the challenge of sampling from complex EBMs. One possible reason is that there are **critical temperatures** around which the temperature transition must be very slow (as the temperature is gradually reduced) in order for tempering to be effective.