

equivalent to minimizing the expected value of

$$\tilde{L}(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^n \left( \frac{\partial^2}{\partial x_j^2} \log p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}) + \frac{1}{2} \left( \frac{\partial}{\partial x_j} \log p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}) \right)^2 \right) \quad (18.25)$$

where  $n$  is the dimensionality of  $\mathbf{x}$ .

Because score matching requires taking derivatives with respect to  $\mathbf{x}$ , it is not applicable to models of discrete data. However, the latent variables in the model may be discrete.

Like the pseudolikelihood, score matching only works when we are able to evaluate  $\log \tilde{p}(\mathbf{x})$  and its derivatives directly. It is not compatible with methods that only provide a lower bound on  $\log \tilde{p}(\mathbf{x})$ , because score matching requires the derivatives and second derivatives of  $\log \tilde{p}(\mathbf{x})$  and a lower bound conveys no information about its derivatives. This means that score matching cannot be applied to estimating models with complicated interactions between the hidden units, such as sparse coding models or deep Boltzmann machines. While score matching can be used to pretrain the first hidden layer of a larger model, it has not been applied as a pretraining strategy for the deeper layers of a larger model. This is probably because the hidden layers of such models usually contain some discrete variables.

While score matching does not explicitly have a negative phase, it can be viewed as a version of contrastive divergence using a specific kind of Markov chain (Hyvärinen, 2007a). The Markov chain in this case is not Gibbs sampling, but rather a different approach that makes local moves guided by the gradient. Score matching is equivalent to CD with this type of Markov chain when the size of the local moves approaches zero.

Lyu (2009) generalized score matching to the discrete case (but made an error in their derivation that was corrected by Marlin *et al.* (2010)). Marlin *et al.* (2010) found that **generalized score matching** (GSM) does not work in high dimensional discrete spaces where the observed probability of many events is 0.

A more successful approach to extending the basic ideas of score matching to discrete data is **ratio matching** (Hyvärinen, 2007b). Ratio matching applies specifically to binary data. Ratio matching consists of minimizing the average over examples of the following objective function:

$$L^{(\text{RM})}(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^n \left( \frac{1}{1 + \frac{p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})}{p_{\text{model}}(f(\mathbf{x}), j; \boldsymbol{\theta})}} \right)^2, \quad (18.26)$$