CSC2535: 2011
Lecture 5a

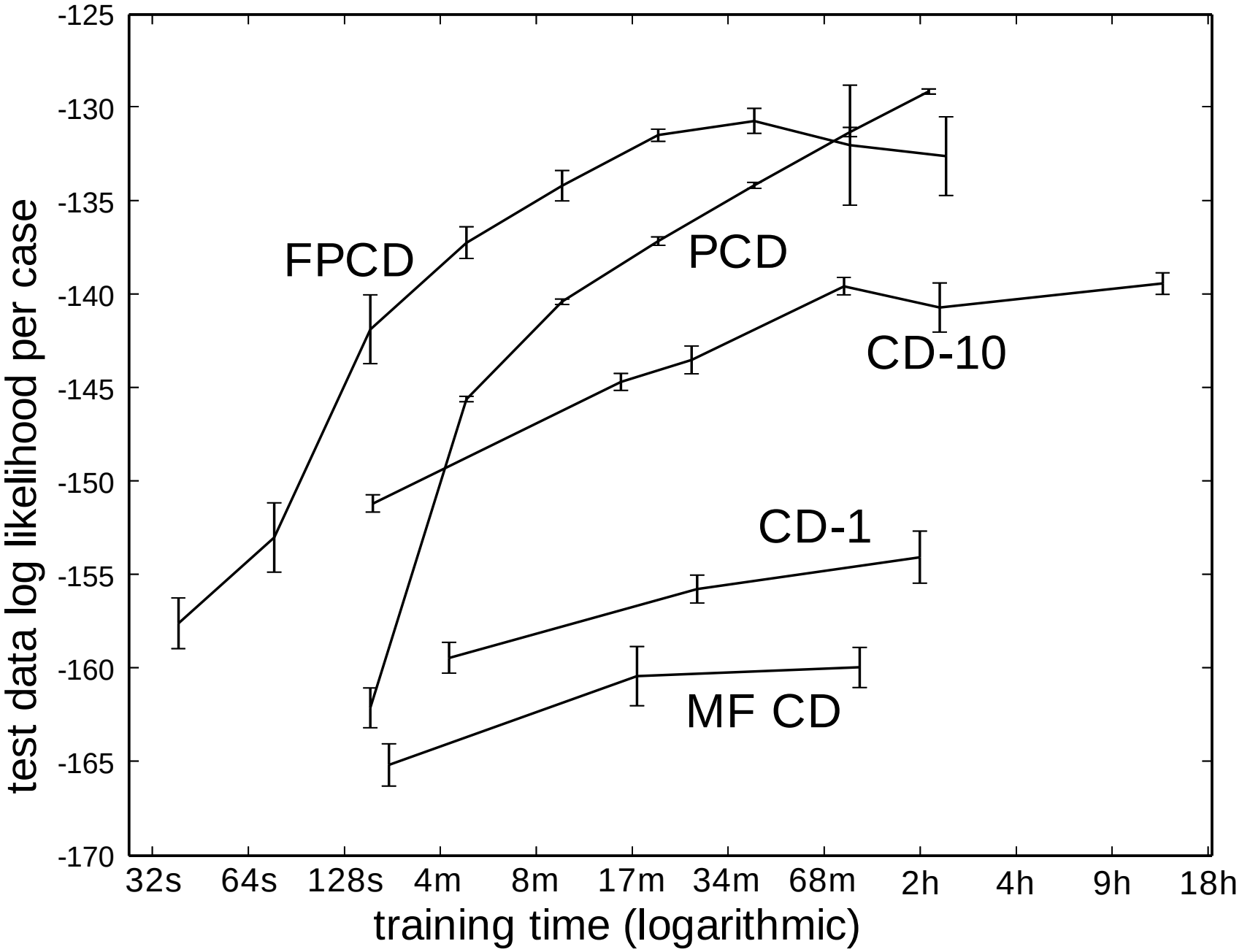
More ways to fit energy-based models

Geoffrey Hinton

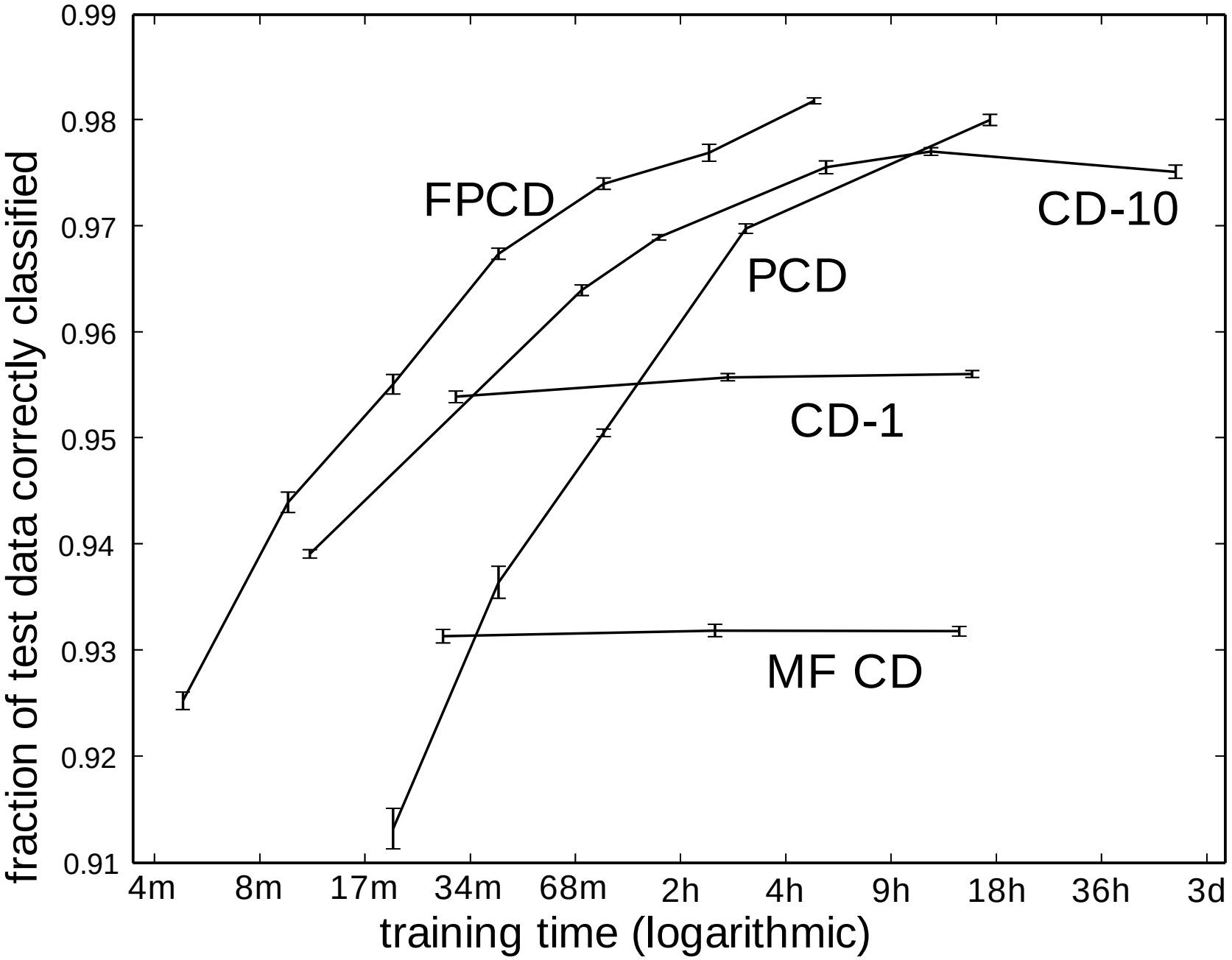
Fast PCD

- Add an extra set of fast weights to make it easier for the energy surface to help the mixing.
  - The fast weights are only used for updating the fantasy particles. They are not part of the model.
- Fast weights learn quickly, but they also have very strong weight decay, so they forget quickly.
  - They act as a thin additive overlay on the energy surface defined by the slow weights.
- Towards the end of learning, we turn down the learning rate for the slow weights, but keep it high for the fast weights.
  - This allows persistent CD to keep mixing well.

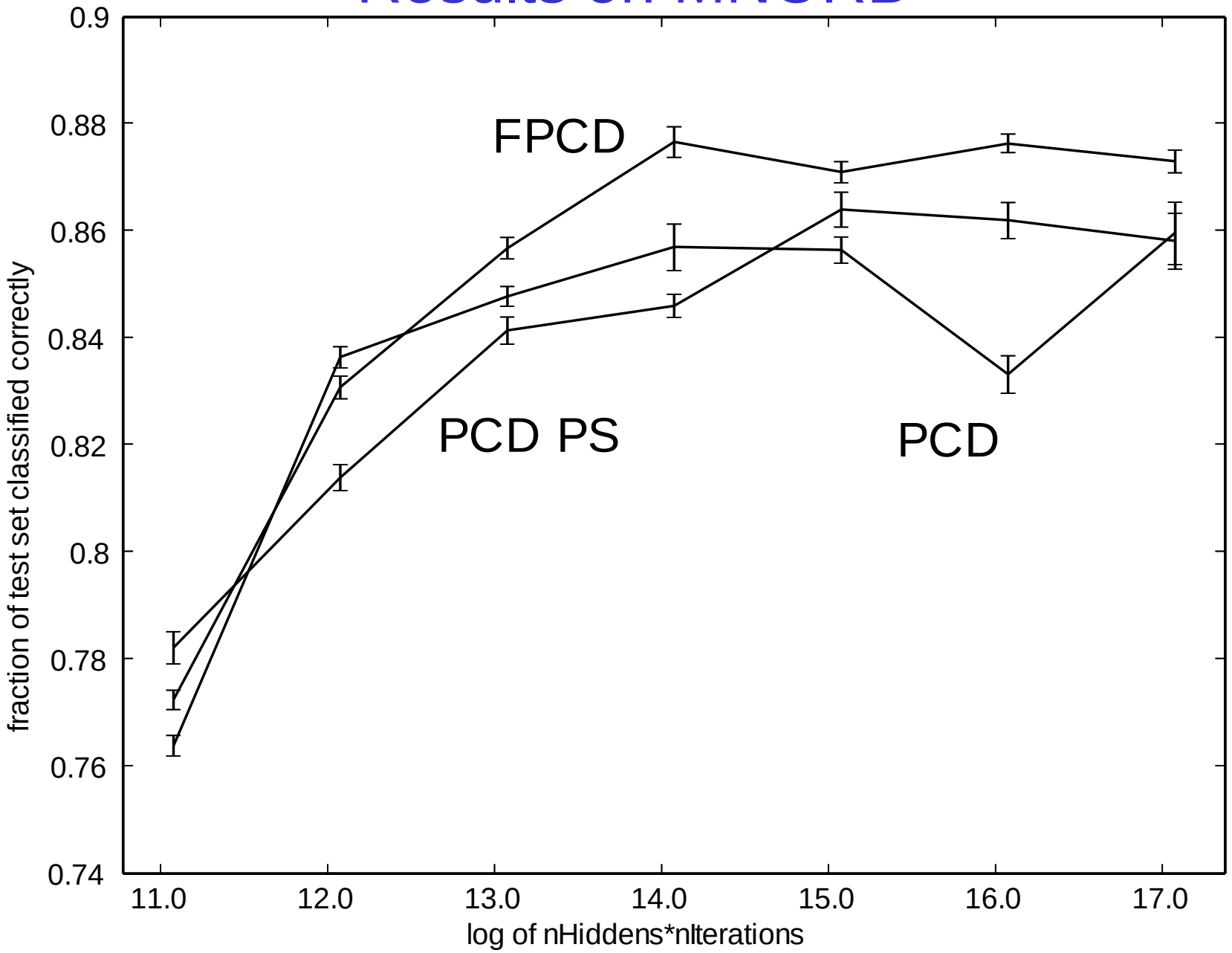
Log density on MNIST with 25 hidden units



Discrimination on MNIST with 500 hidden units



Results on MNORB



Pseudo-likelihood

- Maximum likelihood maximizes:

$$\sum_{i=1}^{n} \log p(\mathbf{x}_i; \theta)$$

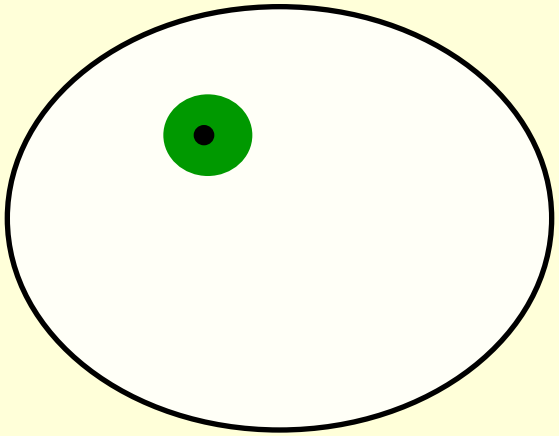
- Pseudo-likelihood maximizes:

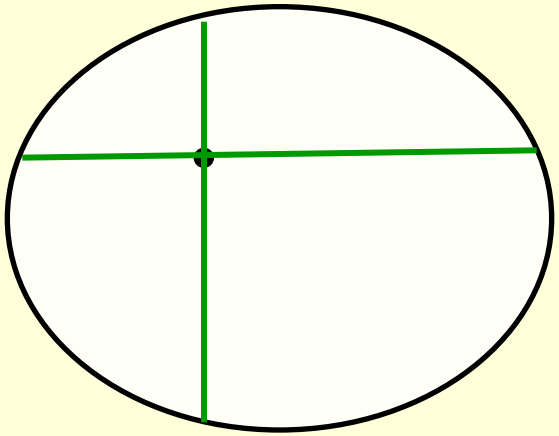
$$\sum_{i=1}^{n} \sum_{d} \log p(x_i^{(d)}; \theta)$$

where *d* is an index over dimensions of the data vectors

Comparison of pseudo-likelihood with CD

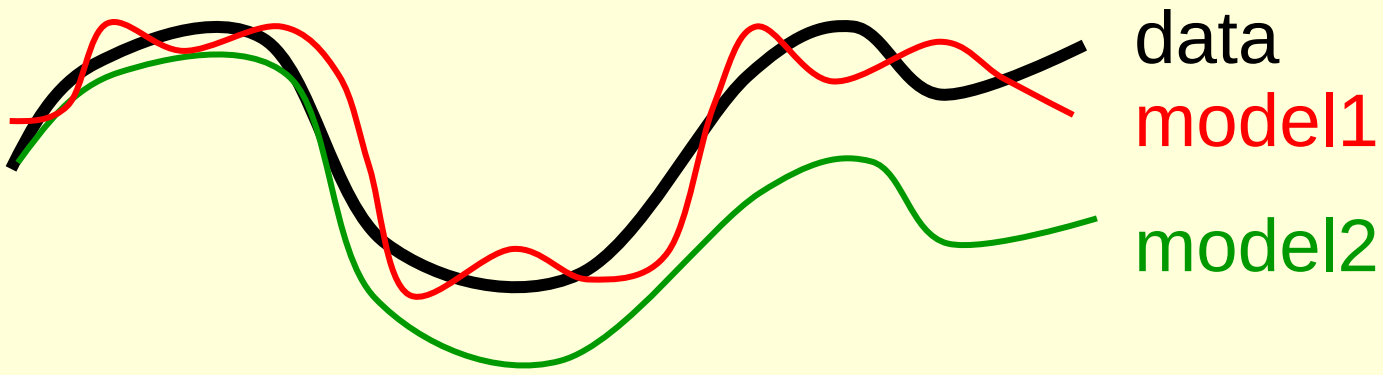
- CD explores the full-dimensional dataspace in the vicinity of each datapoint.
- Pseudo-likelihood explores the whole of each possible axis-aligned one-dimensional space.
  - This has its own problems!





Score matching

- For energy-based models, the gradient of the unnormalized log prob is usually tractable.
  - So instead of matching the log prob of the model to the empirical distribution, match the gradient of the log prob of the model to the gradient of the log prob of the empirical distribution.



Which model is best? Which is best for denoising?

Computing the integral of the squared difference of the model and data gradients

- It looks as if we need to fit a non-parametric model to the data in order to estimate the gradient of the empirical distribution. But there is a neat trick:
- We can use the gradient and curvature of the log prob that the model assigns to the datapoints to estimate the integral of the squared difference in gradients.

$$\frac{1}{2}\int_{\xi\in R^n} p_x(\xi)\left\|\psi(\xi;\theta)-\psi_x(\xi)\right\|^2 d\xi$$

$$=\int_{\xi\in R^n} p_x(\xi)\sum_{i=1}^{n}\left[\partial_i\psi_i(\xi;\theta)+\frac{1}{2}\psi_i(\xi;\theta)^2\right]d\xi+\text{const}$$

When CD becomes score matching

- Suppose we do CD on continuous data with a Langevin Markov chain that works as follows:
- Choose a momentum vector from a Gaussian distribution and take one step in the direction of the gradient plus the added noise. Use a step size that is scaled by the amount of added noise.
- In the limit, as the noise variance goes to zero, this version of CD is the same as score matching.
  - Proved by Hyvarinen.