Minimum Probability Flow Learning

Jascha Sohl-Dickstein

Peter Battaglino

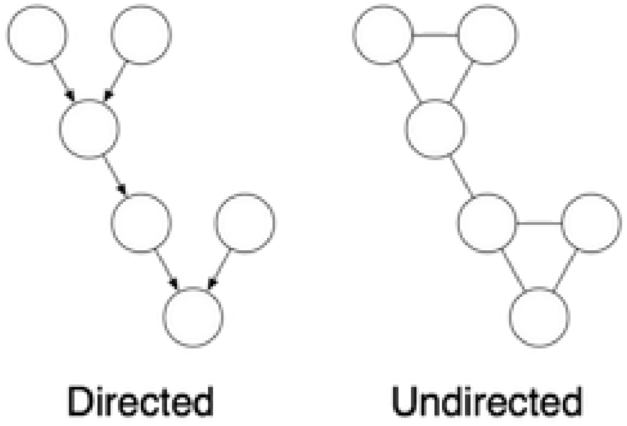
Michael R. DeWeese

U.C. Berkeley

ICML 2011 Distinguished Paper Award

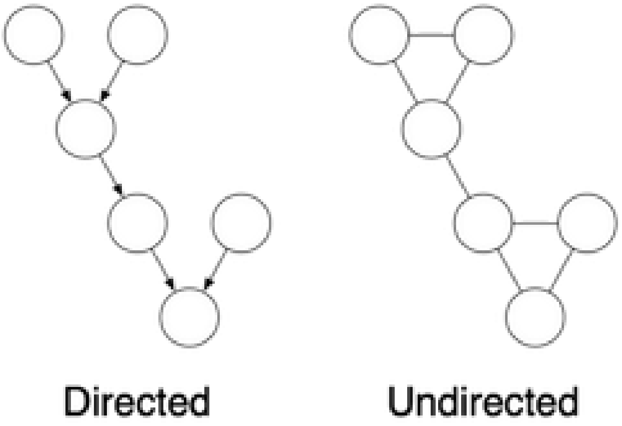
Background: Graphical Model

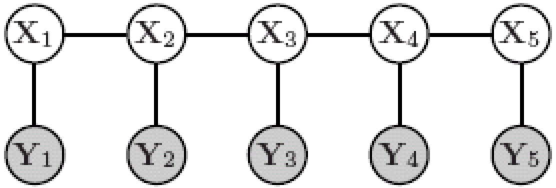
- Graphical model
  - undirected graph
    - Markov random field (MRF)
    - Conditional random field (CRF)
  - directed graph
    - Bayesian network
    - LDA ..



Background: Graphical Model

- Graphical model
  - undirected graph
    - Markov random field (MRF)
    - Conditional random field (CRF)
  - directed graph
    - Bayesian network
    - LDA ..





In this talk, we don't consider hidden variables

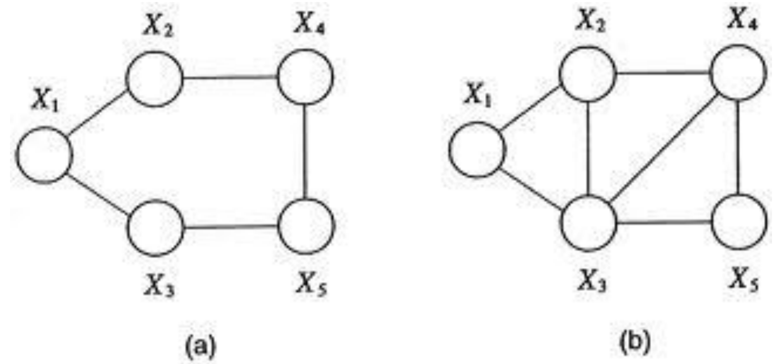
Markov Random Field

- $\boldsymbol{X}=x_1 \ldots x_n$ are n binary random variables

$$p(\mathbf{X}) = \frac{1}{Z}\exp\left[\sum_{ij} J_{ij}x_i x_j + \sum_i h_i x_i\right]$$

$$Z(J,h) = \sum_{\mathbf{x}} \exp\left[\sum_{ij} J_{ij}x_i x_j + \sum_i h_i x_i\right]$$

  - Z is normalization constant
  - aka. “partition function”
  - intractable to evaluate

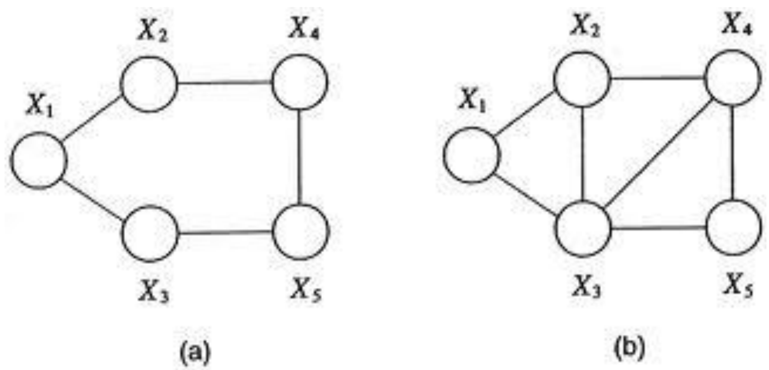


Markov Random Field

- ***X***=$x_1 \ldots x_n$ are *n* binary random variables

$$p(\mathbf{X}) = \frac{1}{Z}\exp\left[\sum_{ij} J_{ij}x_i x_j + \sum_i h_i x_i\right]$$

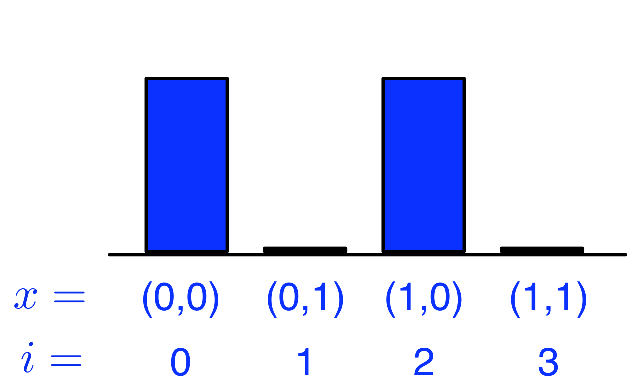
$$Z(J,h) = \sum_{\mathbf{x}} \exp\left[\sum_{ij} J_{ij}x_i x_j + \sum_i h_i x_i\right]$$



- Tasks
  - Inference
    - [given params] for any data ***X***, calculate p(***X***)
  - Learning
    - [given data $\boldsymbol{X}_1..\boldsymbol{X}_d$] learn the parameters J, h

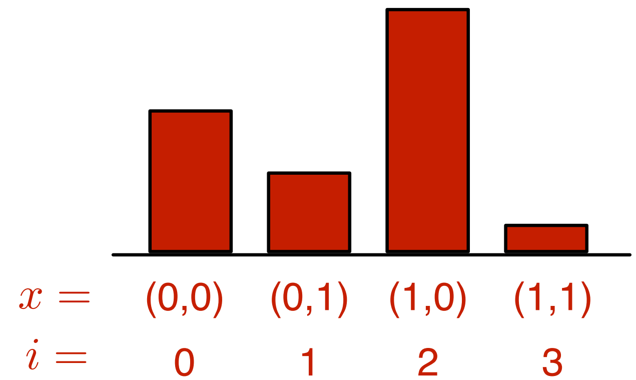
Learning in probabilistic models...

- Want to fit a parametric model to data



data distribution

$$p_i^{(0)} = \text{fraction data in state } i$$



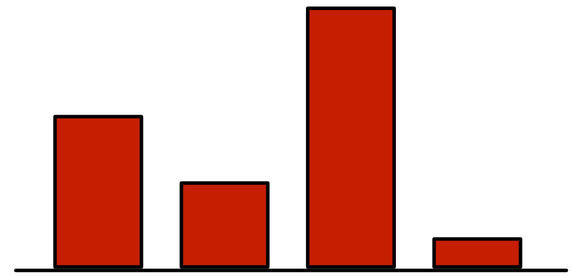
model distribution

$$p_i^{(\infty)}(\theta) = \frac{e^{-E_i(\theta)}}{Z(\theta)}$$

$$Z(\theta) = \sum_i e^{-E_i(\theta)}$$

- Adjust θ so the model distribution looks like the data distribution

Learning in probabilistic models... ...is hard



$$p_i^{(\infty)}(\theta) = \frac{e^{-E_i(\theta)}}{Z(\theta)}$$

$$Z(\theta) = \sum_i e^{-E_i(\theta)}$$

- Maximum likelihood

$$\begin{aligned} K_{ML} &= -\sum_i p_i^{(0)} \log p_i^{(\infty)}(\theta) \\ &= \sum_i p_i^{(0)} E_i(\theta) + \log Z(\theta) \end{aligned}$$

- For a 100 bit binary system

$$Z(\theta) = \sum_{i=1}^{2^{100}} e^{-E_i(\theta)}$$

$$2^{100} = 1267650600228229401496703205376$$

Existing Techniques

- Numerical integration, Monte Carlo sampling, mean field theory, variational bayes, pseudo likelihood, Ratio Matching, Noise Contrastive Estimation...

- Contrastive Divergence

GE Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation* (2002)

- Score Matching

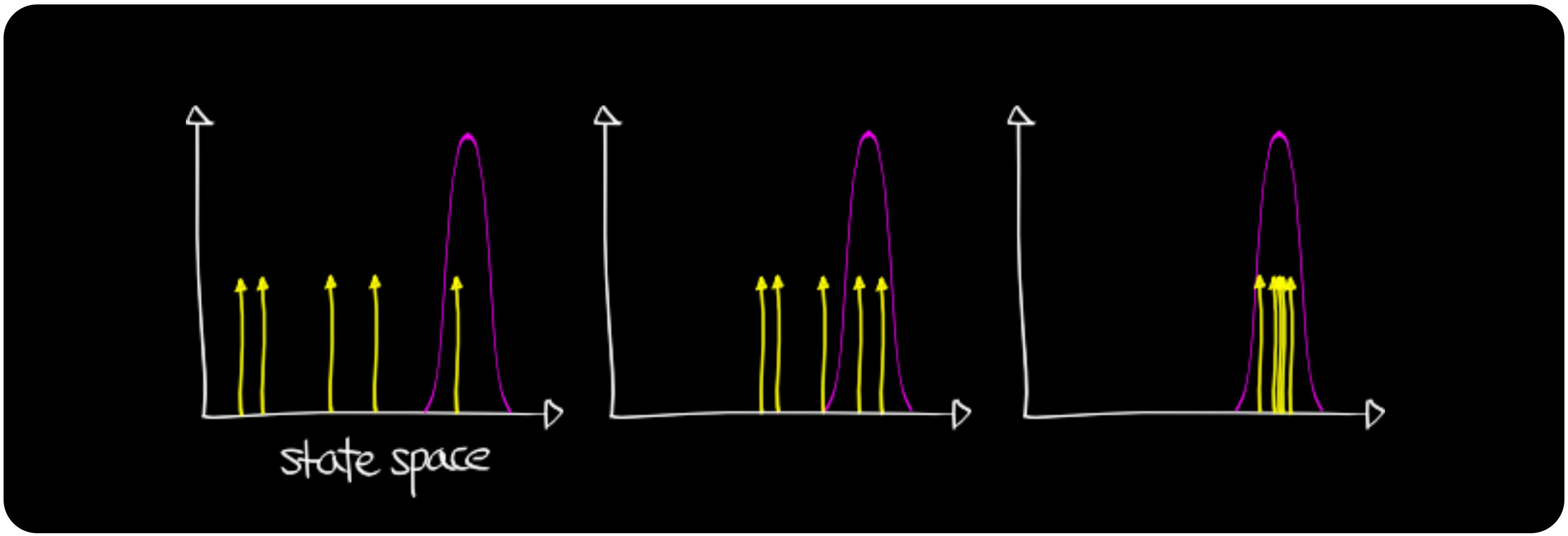
A Hyvärinen. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

- Minimum Velocity learning

J R Movellan. A minimum velocity approach to learning. *unpublished draft*, Jan 2008.

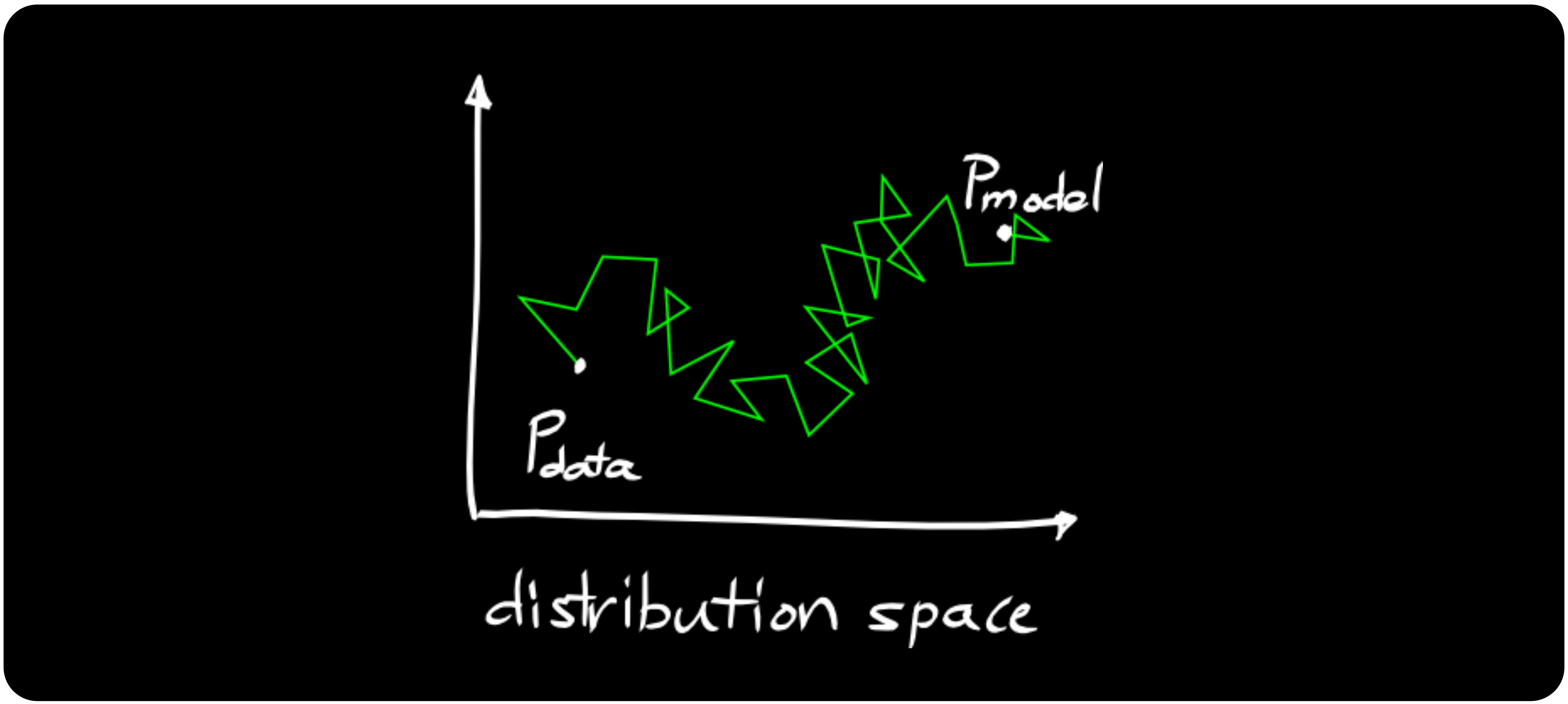
MPF Overview

- Sampling from a distribution:
  - Take a set of samples and apply a series of stochastic transformations to it until it looks like it came from the model distribution



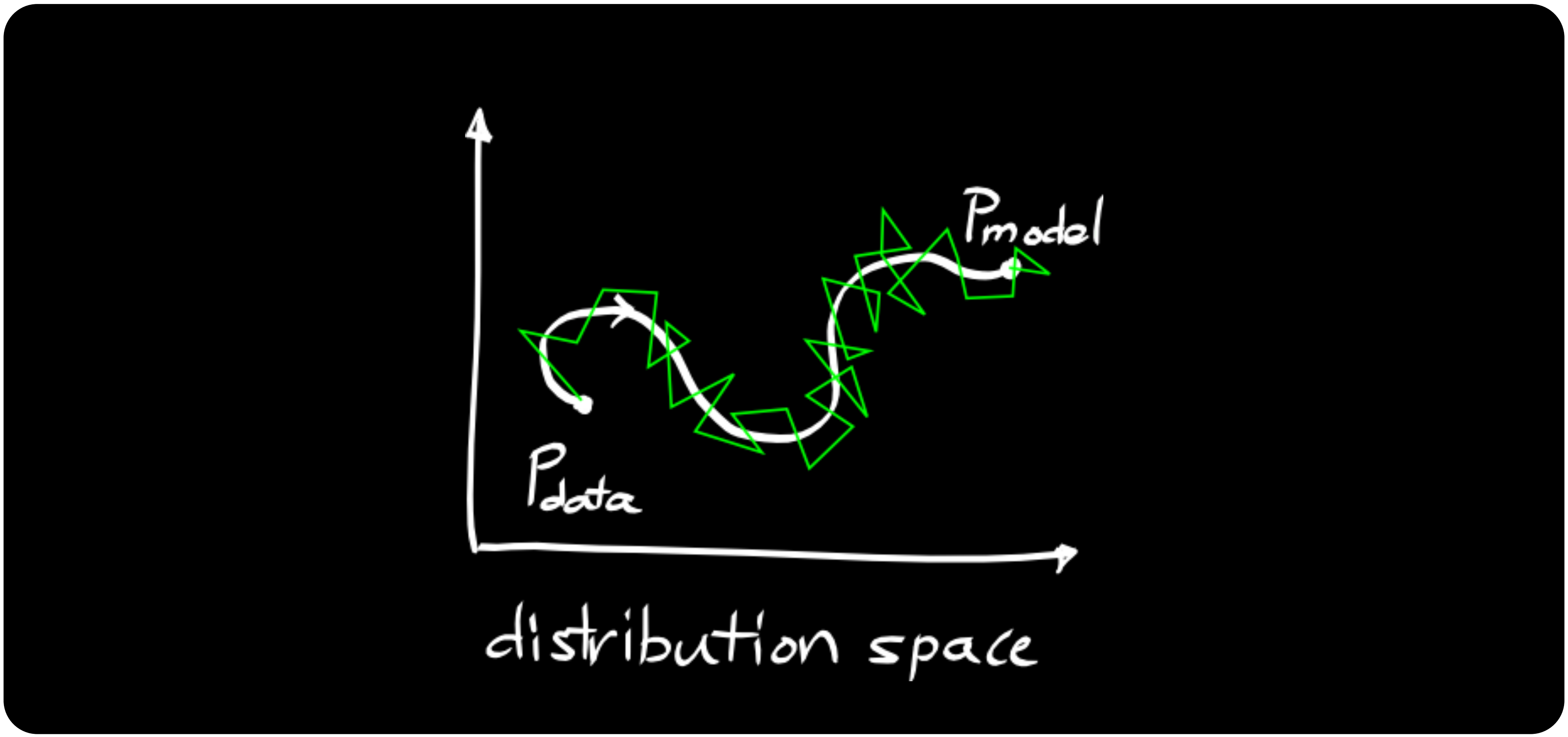
MPF Overview

- Problem with sampling:
  - SLOW to converge for large, high-dimensional data sets



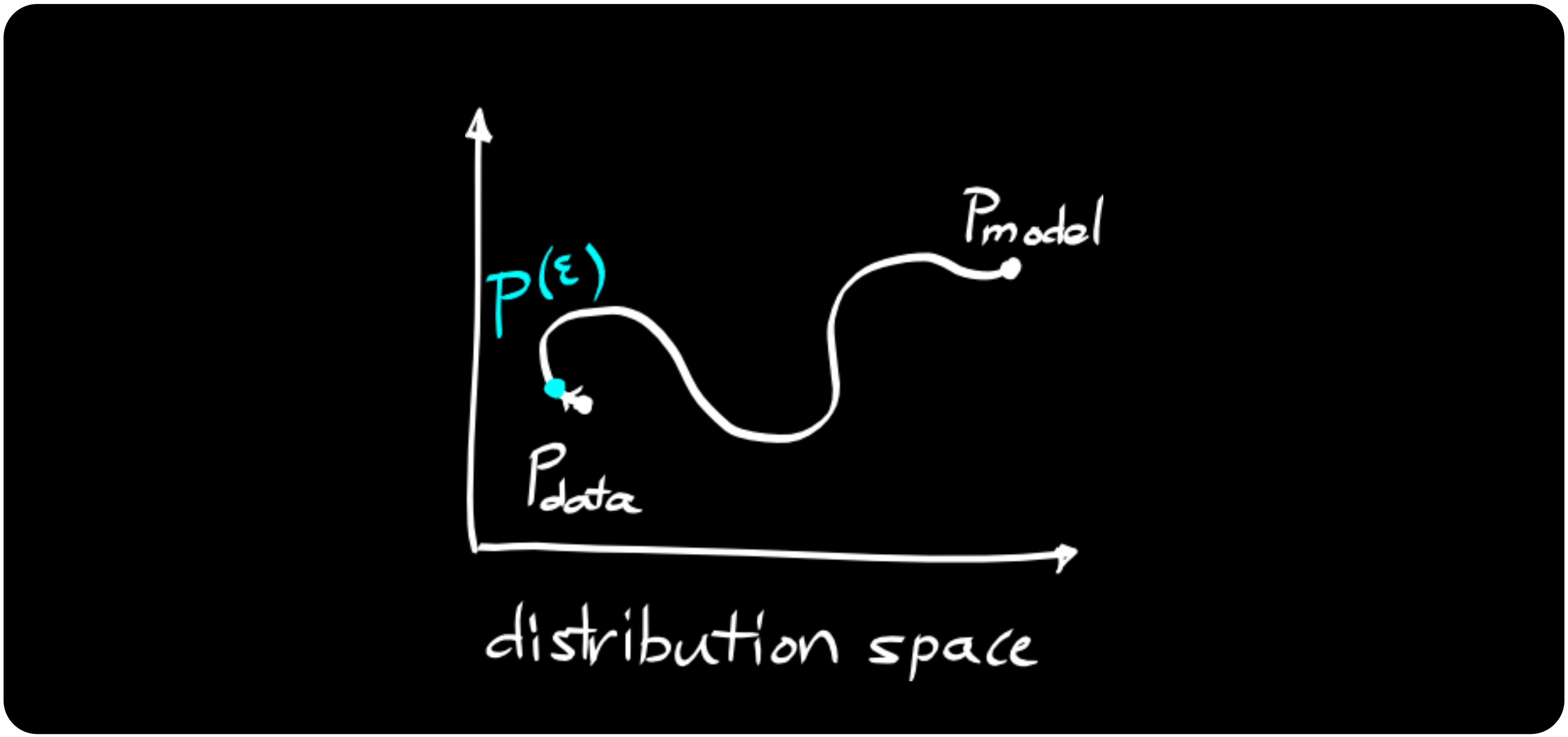
MPF Overview

- Idea: introduce deterministic dynamics interpolating between the data and model distributions...

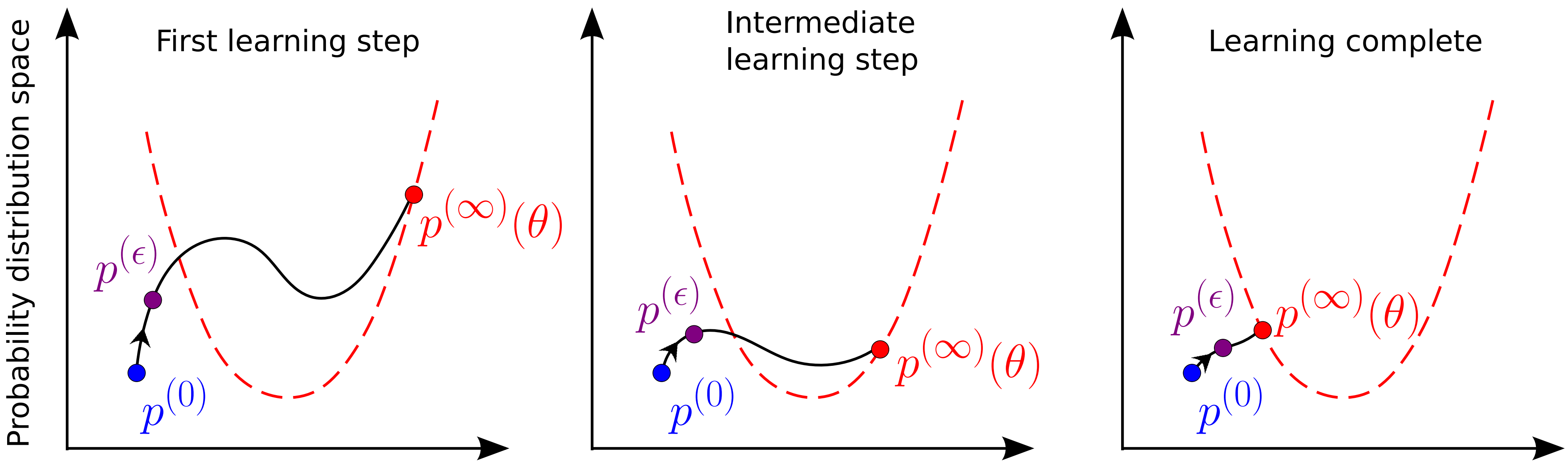


MPF Overview

- ...and only compare the data distribution to the distribution obtained by evolving the dynamics for a small time ε!



Minimum probability flow
Overview



Master Equation

- Transition rates Γ_{ij}
- Master equation conserves probability

$$\dot{p}_i^{(t)} = \sum_{j \neq i} \Gamma_{ij}(\theta)\, p_j^{(t)} - \sum_{j \neq i} \Gamma_{ji}(\theta)\, p_i^{(t)}$$

flow into state i from other states j

flow into other states j from state i

- or in matrix form...:

$$\Gamma_{ii} := -\sum_{j \neq i} \Gamma_{ji}$$

$$\dot{\mathbf{p}}^{(t)} = \boldsymbol{\Gamma} \mathbf{p}^{(t)}$$

$$\mathbf{p}^{(t)} = \exp\left(\boldsymbol{\Gamma} t\right) \mathbf{p}^{(0)}$$

Detailed Balance

- Detailed balance

$$\Gamma_{ji}\ p_i^{(\infty)}(\theta) = \Gamma_{ij}\ p_j^{(\infty)}(\theta)$$

- Choose Γ to converge to model distribution

$$\frac{\Gamma_{ij}}{\Gamma_{ji}} = \frac{p_i^{(\infty)}(\theta)}{p_j^{(\infty)}(\theta)} = \exp\left[E_j(\theta) - E_i(\theta)\right]$$

$$\Gamma_{ij} = g_{ij} \exp\left[\frac{1}{2}\left(E_j(\theta) - E_i(\theta)\right)\right]$$

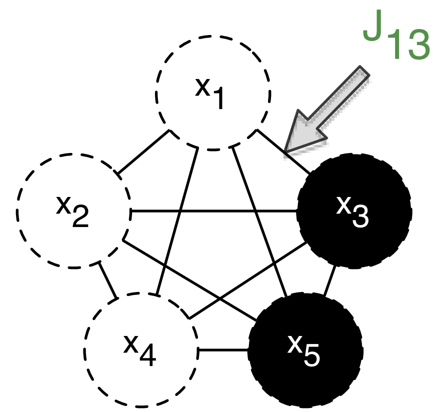
$$g_{ij} = g_{ji} = \begin{cases} 0 & \text{unconnected states} \\ 1 & \text{connected states} \end{cases}$$

Demo Code

- 6 unit Ising model

$$p^{(\infty)}(\mathbf{x}; \mathbf{J}) = \frac{1}{Z(\mathbf{J})} \exp\left[-\sum_{i,j} J_{ij} x_i x_j\right] \qquad \mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$$x_i \in \{0, 1\}$$



- 2 dimensional random projection of $\mathbf{p}^{(t)}$

- $\mathbf{p}^{(0)}$ 150 samples using random J
- $\mathbf{p}^{(\infty)}(\theta)$ initialized to another random J

Objective Function

- Minimize $D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\epsilon)}(\theta)\right)$, for small ϵ

$$\hat{\theta} = \arg\min_{\theta} K_{MPF}(\theta)$$

$$\begin{aligned} K_{MPF}(\theta) = D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\epsilon)}(\theta)\right) &\approx D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\mathbf{t})}(\theta)\right)\Big|_{t=0} + \epsilon \frac{\partial D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\mathbf{t})}(\theta)\right)}{\partial t}\Big|_{t=0} \\ &= \epsilon \sum_{j \notin \text{data}} \dot{p}_j^{(0)} \\ &= \epsilon \sum_{i \notin \text{data}} \sum_{j \in \text{data}} g_{ij} \exp\left[\frac{1}{2}\left(E_j(\theta) - E_i(\theta)\right)\right] p_j^{(0)} \end{aligned}$$

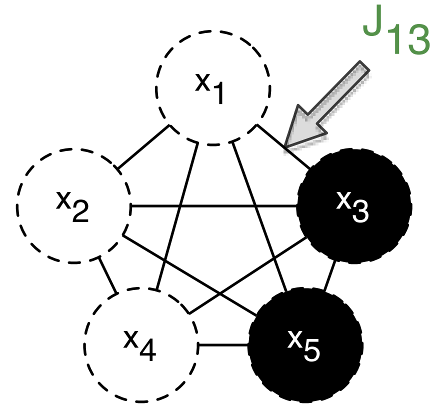
- Minimize initial probability flow from data states to non-data states
- No sampling!

Demo Code

- 6 unit Ising model

$$p^{(\infty)}(\mathbf{x};\mathbf{J}) = \frac{1}{Z(\mathbf{J})} \exp\left[-\sum_{i,j} J_{ij} x_i x_j\right] \qquad \mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$$x_i \in \{0,1\}$$



- 2 dimensional random projection of $\mathbf{p}^{(t)}$

- $\mathbf{p}^{(0)}$ 150 samples using random J
- $\mathbf{p}^{(\infty)}(\theta)$ initialized to another random J

Tractability

- Data distribution $\mathbf{p}^{(0)}$ highly sparse
  - Ignore every column of Γ_{ij} for which $\mathbf{p}_j^{(0)} = 0$
- Γ_{ij} is highly sparse
  - Each state connected to only a small number of other states (eg, within Hamming ball)
- Objective function evaluation costs O(number data points × number connections per data point)

$$K_{MPF}(\theta) = \epsilon \sum_{i \notin \text{data}} \sum_{j \in \text{data}} \Gamma_{ij} p_j^{(0)}$$

Contrastive Divergence

$$\Delta\theta_{CD} \propto -\sum_{i\notin\text{data}}\sum_{j\in\text{data}} p_j^{(0)}\left[\frac{\partial E_j(\theta)}{\partial\theta} - \frac{\partial E_i(\theta)}{\partial\theta}\right] \text{[probability of MCMC step from j} \rightarrow \text{i]}$$

$$\frac{\partial K_{MPF}(\theta)}{\partial\theta} = \epsilon\sum_{i\notin\text{data}}\sum_{j\in\text{data}} p_j^{(0)}\left[\frac{\partial E_j(\theta)}{\partial\theta} - \frac{\partial E_i(\theta)}{\partial\theta}\right] g_{ij}\exp\left[\frac{1}{2}\left(E_j(\theta) - E_i(\theta)\right)\right]$$

- Markov Chain sampling/rejection step replaced by weighting factor
- Objective function!
- Unique global minima when model and data agree

Continuous State Spaces

- Analogous to sum → integral transition

$$p_i^{(0)} = \begin{array}{c}\text{fraction data}\\ \mathcal{D} \text{ in state } i\end{array}$$

$$p^{(0)}(\mathbf{x}) = \frac{1}{\mathcal{D}} \sum_{\mathbf{x}_m \in \mathcal{D}} \delta(\mathbf{x} - \mathbf{x}_m)$$

$$p_i^{(t)}$$

$$p^{(t)}(\mathbf{x})$$

$$p_i^{(\infty)}(\theta) = \frac{\exp\left[-E_i(\theta)\right]}{Z(\theta)}$$

$$p^{(\infty)}(\mathbf{x}; \theta) = \frac{\exp\left[-E(\mathbf{x}; \theta)\right]}{Z(\theta)}$$

$$\Gamma_{ij} = g_{ij} \exp\left[\frac{1}{2}\left(E_j(\theta) - E_i(\theta)\right)\right]$$

$$\Gamma(\mathbf{x}_j \rightarrow \mathbf{x}_j) = g(\mathbf{x}_j \rightarrow \mathbf{x}_j) \exp\left[\frac{1}{2}\left(E(\mathbf{x}_j; \theta) - E(\mathbf{x}_i; \theta)\right)\right]$$

Score Matching

$$g\left(\mathbf{x}_j \rightarrow \mathbf{x}_i\right) = g\left(\mathbf{x}_i \rightarrow \mathbf{x}_j\right) = \begin{cases} 1 & \left\|\mathbf{x}_j - \mathbf{x}_i\right\|_2 < r \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \lim_{r \rightarrow 0} K_{MPF} &\propto K_{SM} \\ &= \left\langle \frac{1}{2} \nabla_{\mathbf{x}} E\left(\mathbf{x};\theta\right) \cdot \nabla_{\mathbf{x}} E\left(\mathbf{x};\theta\right) - \nabla^2_{\mathbf{x}} E\left(\mathbf{x};\theta\right) \right\rangle_{p^{(0)}(\mathbf{x})} \end{aligned}$$

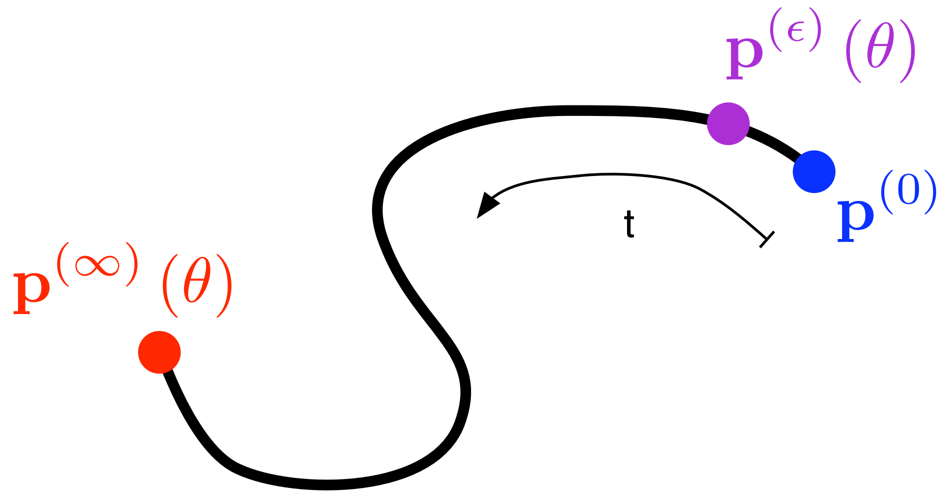
Objective Functions

- Maximum Likelihood

$$K_{ML} = D_{KL}\left(\mathbf{p}^{(\mathbf{0})} || \mathbf{p}^{(\infty)}\left(\theta\right)\right)$$

- Minimum Probability flow

$$K_{MPF} = D_{KL}\left(\mathbf{p}^{(\mathbf{0})} || \mathbf{p}^{(\epsilon)}\left(\theta\right)\right)$$



- Contrastive Divergence

$$K_{CD} \approx D_{KL}\left(\mathbf{p}^{(\mathbf{0})} || \mathbf{p}^{(\infty)}\left(\theta\right)\right) - D_{KL}\left(\mathbf{p}^{(\mathbf{1})}\left(\theta\right) || \mathbf{p}^{(\infty)}\left(\theta\right)\right)$$

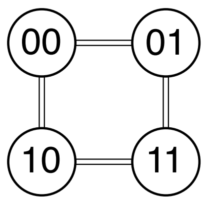
- Score Matching

$$K_{SM} = \left\langle \frac{1}{2} \nabla_{\mathbf{x}} E\left(\mathbf{x};\theta\right) \cdot \nabla_{\mathbf{x}} E\left(\mathbf{x};\theta\right) - \nabla^2_{\mathbf{x}} E\left(\mathbf{x};\theta\right) \right\rangle_{p^{(0)}(\mathbf{x})}$$

Connectivity

- Discrete space
  - Nearest neighbors

$$g_{ij} = g_{ji} = \begin{cases} 1 & i,j \text{ differ by 1 bit flip} \\ 0 & \text{otherwise} \end{cases}$$



Connectivity

- Continuous space
  - Hamiltonian dynamics (similar to hybrid Monte Carlo)

▸ Extend distribution to include auxiliary momentum variables **q**

$$p^{(\infty)}\left(\mathbf{x};\theta\right) \quad \Rightarrow \quad p^{(\infty)}\left(\mathbf{x},\mathbf{q};\theta\right) = p^{(\infty)}\left(\mathbf{x};\theta\right)p^{(\infty)}\left(\mathbf{q}\right) = \frac{e^{-H\left(\mathbf{x},\mathbf{q};\theta\right)}}{Z_H\left(\theta\right)}$$

$$H\left(\mathbf{x},\mathbf{q};\theta\right) = E\left(\mathbf{x};\theta\right) + \frac{1}{2}\left|\left|q\right|\right|_2^2$$

Connectivity

- Continuous space

$$p^{(\infty)}\left(\mathbf{x};\theta\right)$$

$$p^{(\infty)}\left(\mathbf{x},\mathbf{q};\theta\right) = p^{(\infty)}\left(\mathbf{x};\theta\right) p^{(\infty)}\left(\mathbf{q}\right) = \frac{e^{-H(\mathbf{x},\mathbf{q};\theta)}}{Z_H\left(\theta\right)}$$

$$H\left(\mathbf{x},\mathbf{q};\theta\right) = E\left(\mathbf{x};\theta\right) + \frac{1}{2}\left|\left|q\right|\right|_2^2$$

▸ Allow connectivity between momenta, and between states separated by leapfrog dynamics

$$\begin{aligned} g\left(\{\mathbf{x}_j,\mathbf{q}_j\} \rightarrow \{\mathbf{x}_i,\mathbf{q}_i\}\right) &= g\left(\{\mathbf{x}_i,\mathbf{q}_i\} \rightarrow \{\mathbf{x}_j,\mathbf{q}_j\}\right) \\ &= \begin{cases} 1 & \mathbf{x}_i = \mathbf{x}_j \\ 1 & \{\mathbf{x}_i,\mathbf{q}_i\} = \text{leapfrog}\left(\{\mathbf{x}_j,\mathbf{q}_j\};\phi\right) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

(transitions where only **q** changes don't effect objective)

Connectivity

- Continuous space

$$p^{(\infty)}\left(\mathbf{x};\theta\right) \quad \Rightarrow \quad p^{(\infty)}\left(\mathbf{x},\mathbf{q};\theta\right) = p^{(\infty)}\left(\mathbf{x};\theta\right) p^{(\infty)}\left(\mathbf{q}\right) = \frac{e^{-H(\mathbf{x},\mathbf{q};\theta)}}{Z_H\left(\theta\right)}$$

$$H\left(\mathbf{x},\mathbf{q};\theta\right) = E\left(\mathbf{x};\theta\right) + \frac{1}{2}\left|\left|q\right|\right|_2^2$$

$$\begin{aligned} g\left(\{\mathbf{x}_j,\mathbf{q}_j\} \rightarrow \{\mathbf{x}_i,\mathbf{q}_i\}\right) &= g\left(\{\mathbf{x}_i,\mathbf{q}_i\} \rightarrow \{\mathbf{x}_j,\mathbf{q}_j\}\right) \\ &= \begin{cases} 1 & \mathbf{x}_i = \mathbf{x}_j \\ 1 & \{\mathbf{x}_i,\mathbf{q}_i\} = \text{leapfrog}\left(\{\mathbf{x}_j,\mathbf{q}_j\};\phi\right) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

▸ Alternate between updating ϕ and minimizing K_{MPF}

1. Set $\phi = \theta$
2. Set $\theta = \arg\min_{\theta} K_{MPF}\left(\theta;\phi\right)$
3. Repeat

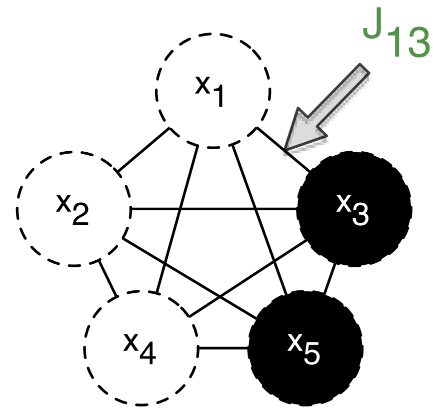
Examples - Ising

- Maximum entropy distribution over binary variables consistent with pairwise statistics

$$p^{(\infty)}(\mathbf{x};\mathbf{J}) = \frac{1}{Z(\mathbf{J})}\exp\left[-\sum_{i,j} J_{ij}x_i x_j\right]$$

$$x_i \in \{0,1\}$$

$$\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$



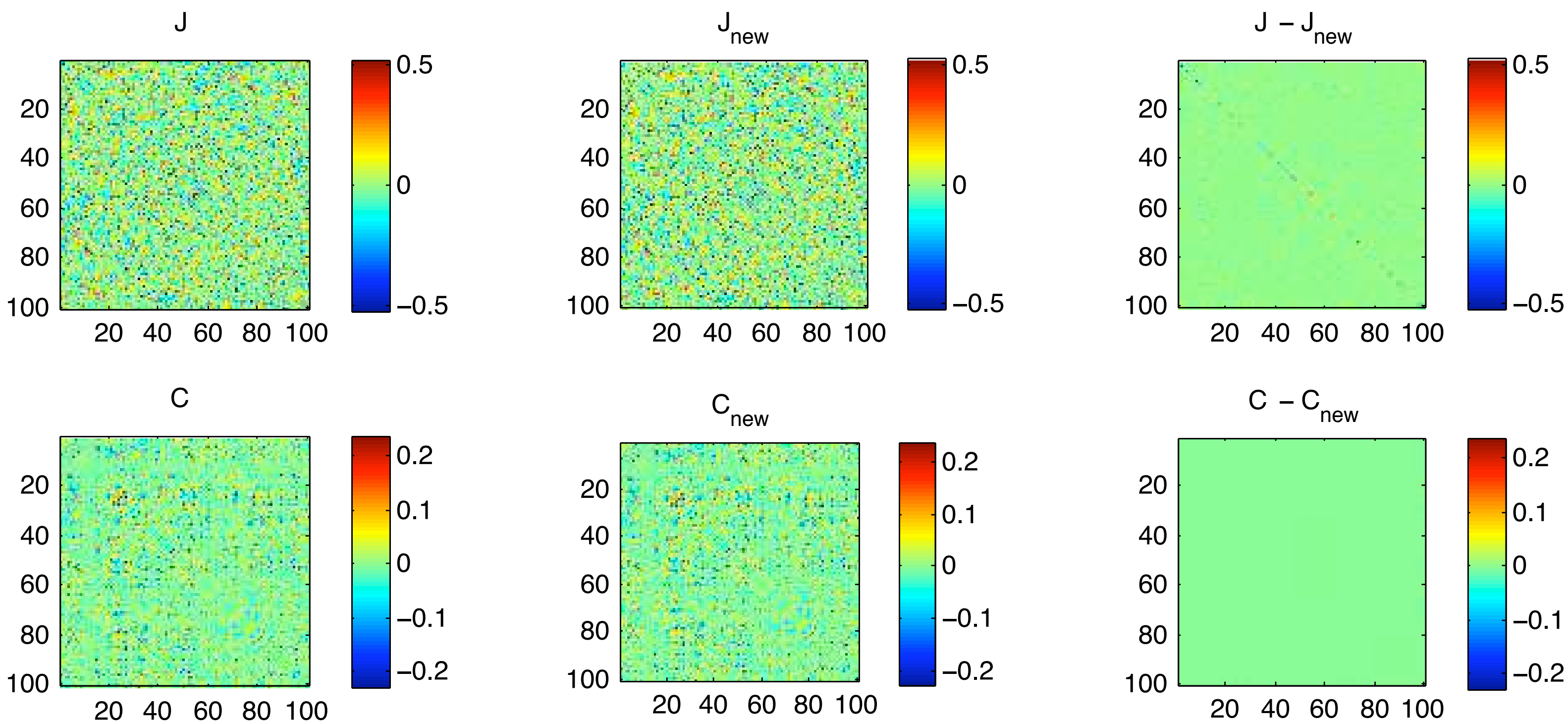
- > 2 orders of magnitude improvement in learning time

T Broderick, M Dudík, G Tkačik, R Schapire, and W Bialek. Faster solutions of the inverse pairwise ising problem. *E-print arXiv*, Jan 2007.

J Shlens, G D Field, J L Gauthier, M Greschner, A Sher, A M Litke, and E J Chichilnisky. The structure of large-scale synchronized firing in primate retina. *Journal of Neuroscience*, 29(15):5022–5031, Apr 2009.

Examples - Ising

- MPF recovers Ising model parameters (100 units, 100,000 samples, J std. dev. 0.04)

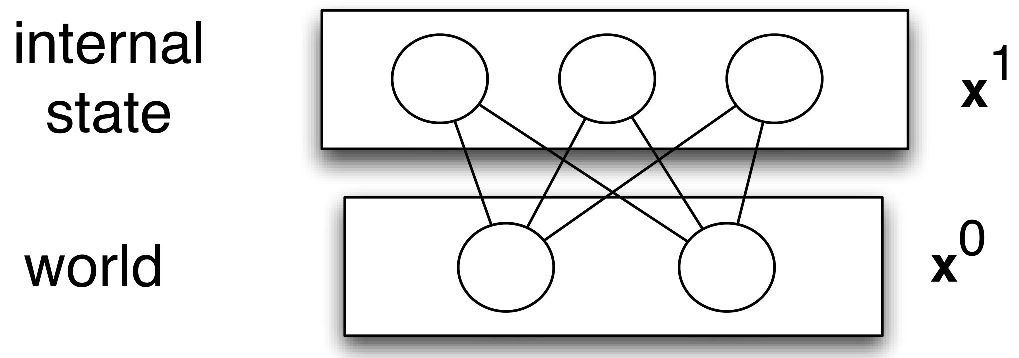


$$p^{(\infty)}(\mathbf{x}; \mathbf{J}) = \frac{1}{Z(\mathbf{J})} \exp\left[-\sum_{i,j} J_{ij} x_i x_j\right]$$

$$x_i \in \{0, 1\}$$

Examples - RBM

- Restricted Boltzmann Machine



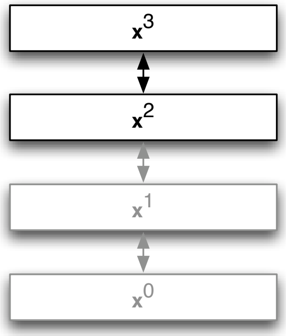
$$E(\mathbf{x}; W) = \mathbf{x}^{0^T} \mathbf{W} \mathbf{x}^1$$

$$p(\mathbf{x}; W) \propto e^{-E(\mathbf{x}; W)}$$

- Explicitly evaluate log likelihood on 20 visible unit, 20 hidden unit RBM
  - random -21.529931 bits
  - MPF -9.044596 bits
  - CD1 -15.822924 bits
  - CD10 -38.011133 bits (!!!) (continuing to increase!)

Examples - DBN

- Deep Belief Network is constructed by stacking RBMs



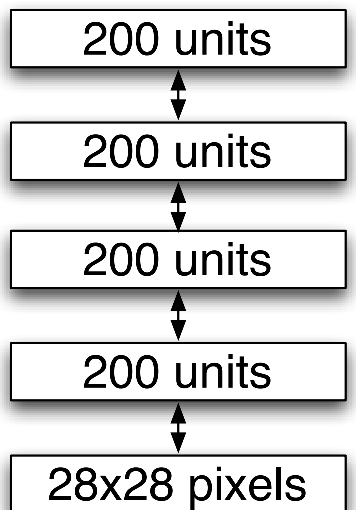
$$E\left(\mathbf{x}^2, \mathbf{x}^3\right) = \mathbf{x}^{2^T} \mathbf{W}^{23} \mathbf{x}^3 \qquad p\left(\mathbf{x}^2, \mathbf{x}^3\right) \propto e^{-E\left(\mathbf{x}^2, \mathbf{x}^3\right)}$$

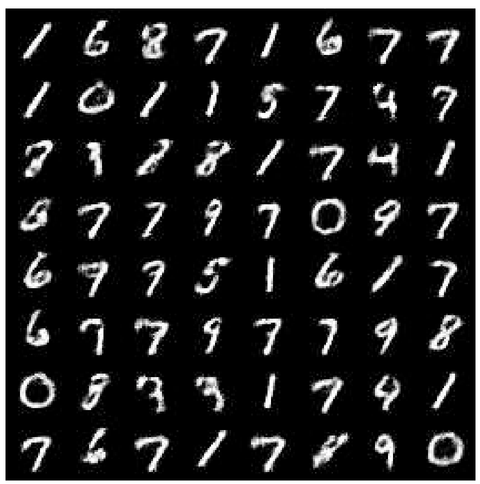
$$E\left(\mathbf{x}^1, \mathbf{x}^2\right) = \mathbf{x}^{1^T} \mathbf{W}^{12} \mathbf{x}^2 \qquad p\left(\mathbf{x}^1, \mathbf{x}^2\right) \propto e^{-E\left(\mathbf{x}^1, \mathbf{x}^2\right)}$$

$$E\left(\mathbf{x}^0, \mathbf{x}^1\right) = \mathbf{x}^{0^T} \mathbf{W}^{01} \mathbf{x}^1 \qquad p\left(\mathbf{x}^0, \mathbf{x}^1\right) \propto e^{-E\left(\mathbf{x}^0, \mathbf{x}^1\right)}$$

$$p\left(\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\right) = p\left(\mathbf{x}^2, \mathbf{x}^3\right) p\left(\mathbf{x}^1 | \mathbf{x}^2\right) p\left(\mathbf{x}^0 | \mathbf{x}^1\right)$$

- Train DBN on MNIST digit database



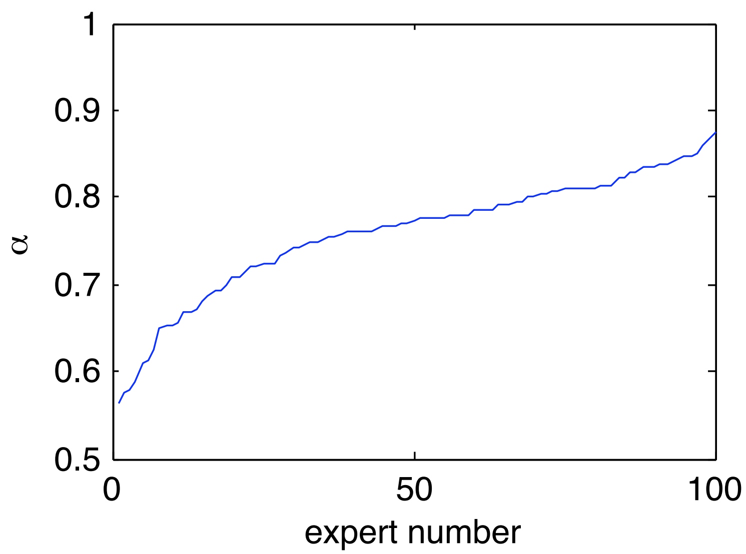


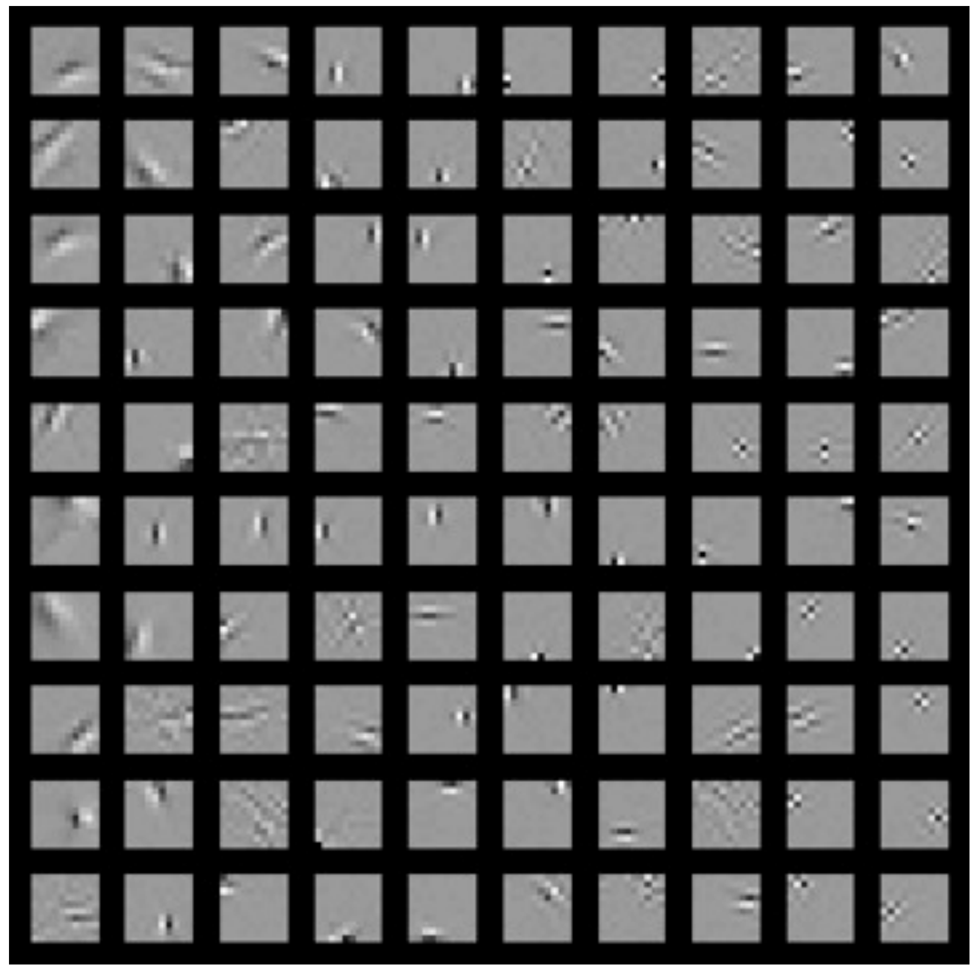
MPF

1 step CD

Examples - Product of Student-t distributions

$$p^{(\infty)}(\mathbf{x}; \mathbf{J}, \alpha) \propto e^{-\sum_i \alpha_i \log\left[1+(\mathbf{J}_i\mathbf{x})^2\right]}$$





MPF Summary

- General method for estimating parameters of probabilistic models
- Well defined objective function, which can be minimized using many known techniques (eg, l-BFGS, minFunc)
- Handles continuous and discrete systems
- Unique global minimum at Maximum Likelihood solution if model can exactly match data
- Convex for **E**(θ) in exponential family (eg Ising model)
- Reduces to Minimum Velocity learning, Score Matching, and (certain forms of) Contrastive Divergence in appropriate limits

Sampling Connectivity

$$\Gamma_{ji}\ p_i^{(\infty)}(\theta) \quad = \quad \Gamma_{ij}\ p_j^{(\infty)}(\theta)$$

$$\langle \Gamma_{ji} \rangle = g_{ji} F_{ji}$$

$$\left\langle \Gamma_{ji}\ p_i^{(\infty)}(\theta) \right\rangle \quad = \quad \left\langle \Gamma_{ij}\ p_j^{(\infty)}(\theta) \right\rangle$$

$$g_{ji} F_{ji}\ p_i^{(\infty)}(\theta) = g_{ij} F_{ij}\ p_j^{(\infty)}(\theta)$$

$$\langle \Gamma_{ji} \rangle\ \ p_i^{(\infty)}(\theta) \quad = \quad \langle \Gamma_{ij} \rangle\ \ p_j^{(\infty)}(\theta)$$

$$\frac{F_{ij}}{F_{ji}} = \frac{g_{ji}}{g_{ij}} \frac{p_i^{(\infty)}(\theta)}{p_j^{(\infty)}(\theta)} = \frac{g_{ji}}{g_{ij}} \exp\left[E_j(\theta) - E_i(\theta)\right]$$

$$F_{ij} = \left(\frac{g_{ji}}{g_{ij}}\right)^{\frac{1}{2}} \exp\left[\frac{1}{2}\left(E_j(\theta) - E_i(\theta)\right)\right]$$

$$r_{ij} \quad \sim \quad \text{rand}\ [0, 1)$$

$$\Gamma_{ij} \quad = \quad \begin{cases} -\sum_{k \neq i} \Gamma_{ki} & i = j \\ F_{ij} & r_{ij} \leq g_{ij} \text{ and } i \neq j \\ 0 & r_{ij} > g_{ij} \text{ and } i \neq j \end{cases}$$

Examples - Power series

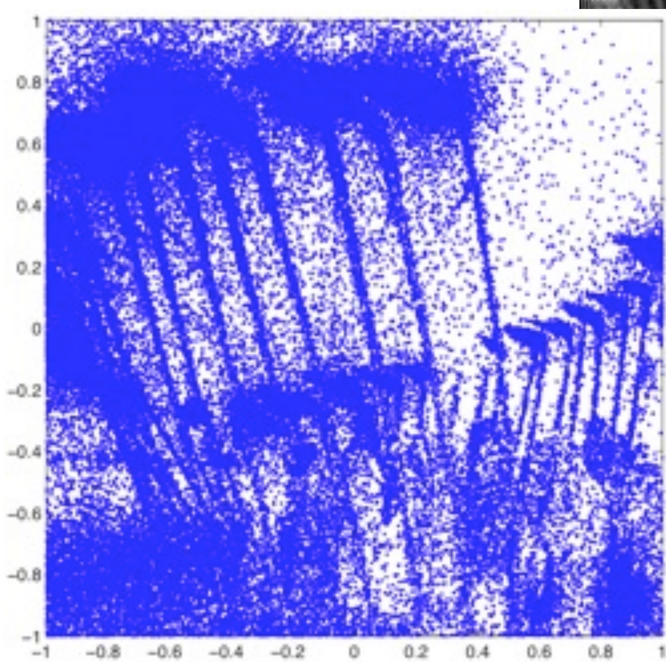
- Fitting a highly unstructured 2-dimensional distribution

$$p^{(\infty)}(x,y;\theta) = \frac{1}{Z(\theta)} \exp\left[-\sum_{m,n=0}^{M} \theta_{mn} L_m(x) L_n(y)\right]$$

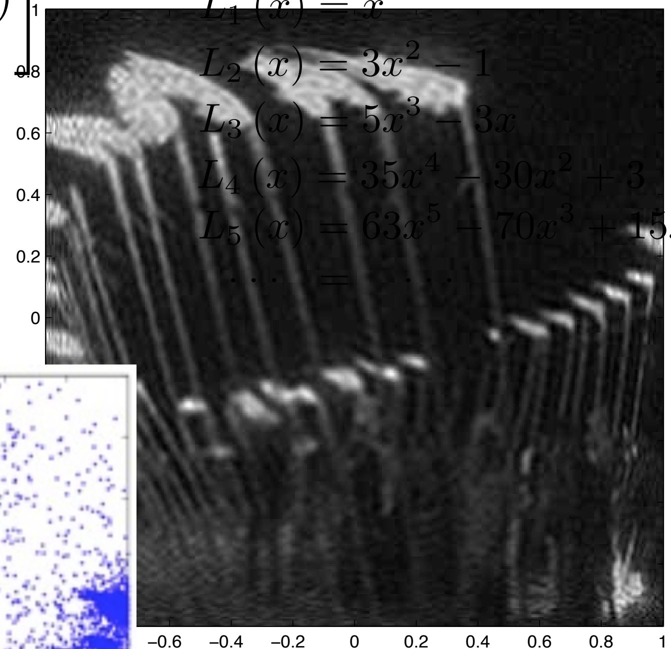
$1]^2$

$L_0(x) = 1$
$L_1(x) = x$
$L_2(x) = 3x^2 - 1$
$L_3(x) = 5x^3 - 3x$
$L_4(x) = 35x^4 - 30x^2 + 3$
$L_5(x) = 63x^5 - 70x^3 + 15x$
$\cdots = \cdots$

data histogram



scatterplot, 100,000 samples

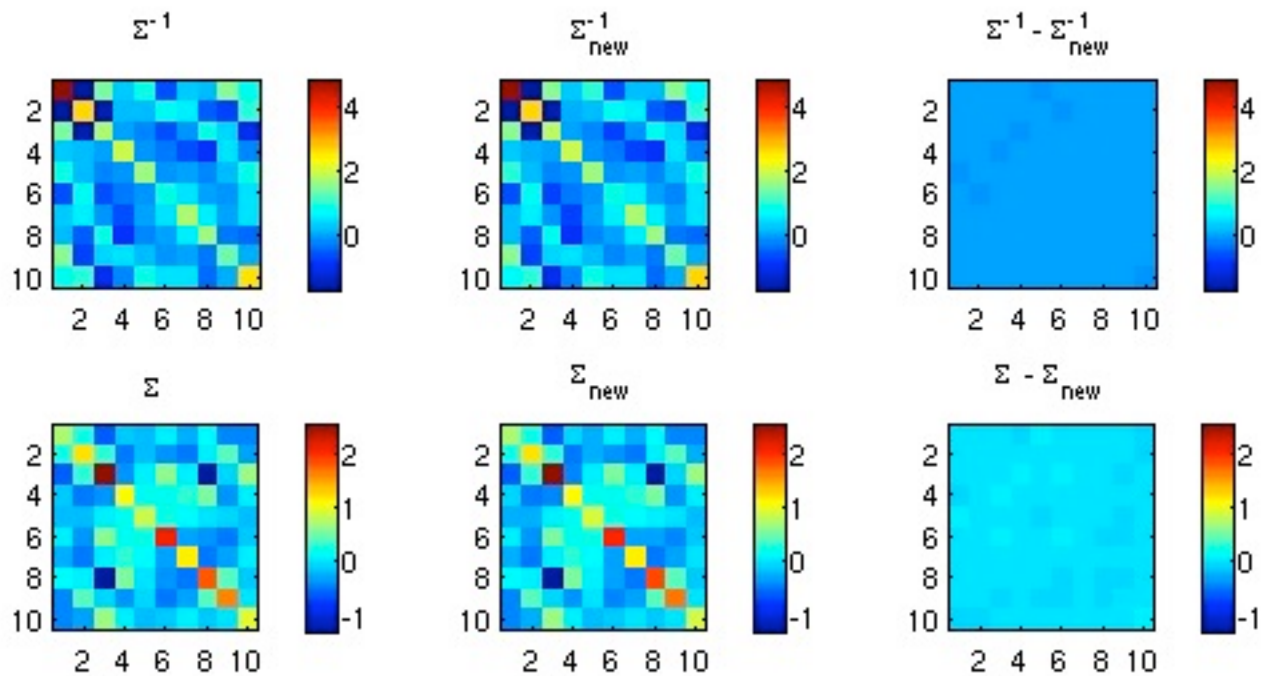


model histogram

Examples - Gaussian

- MPF recovers parameters from 10,000 samples of a 10-dimensional Gaussian distribution

$$p^{(\infty)}(\mathbf{x};\boldsymbol{\Sigma}^{-1}) = \frac{1}{Z\left(\boldsymbol{\Sigma}^{-1}\right)} \exp\left[-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right]$$



Relationship to CD

$$K_{CD} \approx D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\infty)}(\theta)\right) - D_{KL}\left(\mathbf{p}^{(\epsilon)}(\theta)||\mathbf{p}^{(\infty)}(\theta)\right)$$

$$K_{MPF} = D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\epsilon)}(\theta)\right)$$

$$D_{KL}(A||C) \leq D_{KL}(A||B) + D_{KL}(B||C)$$

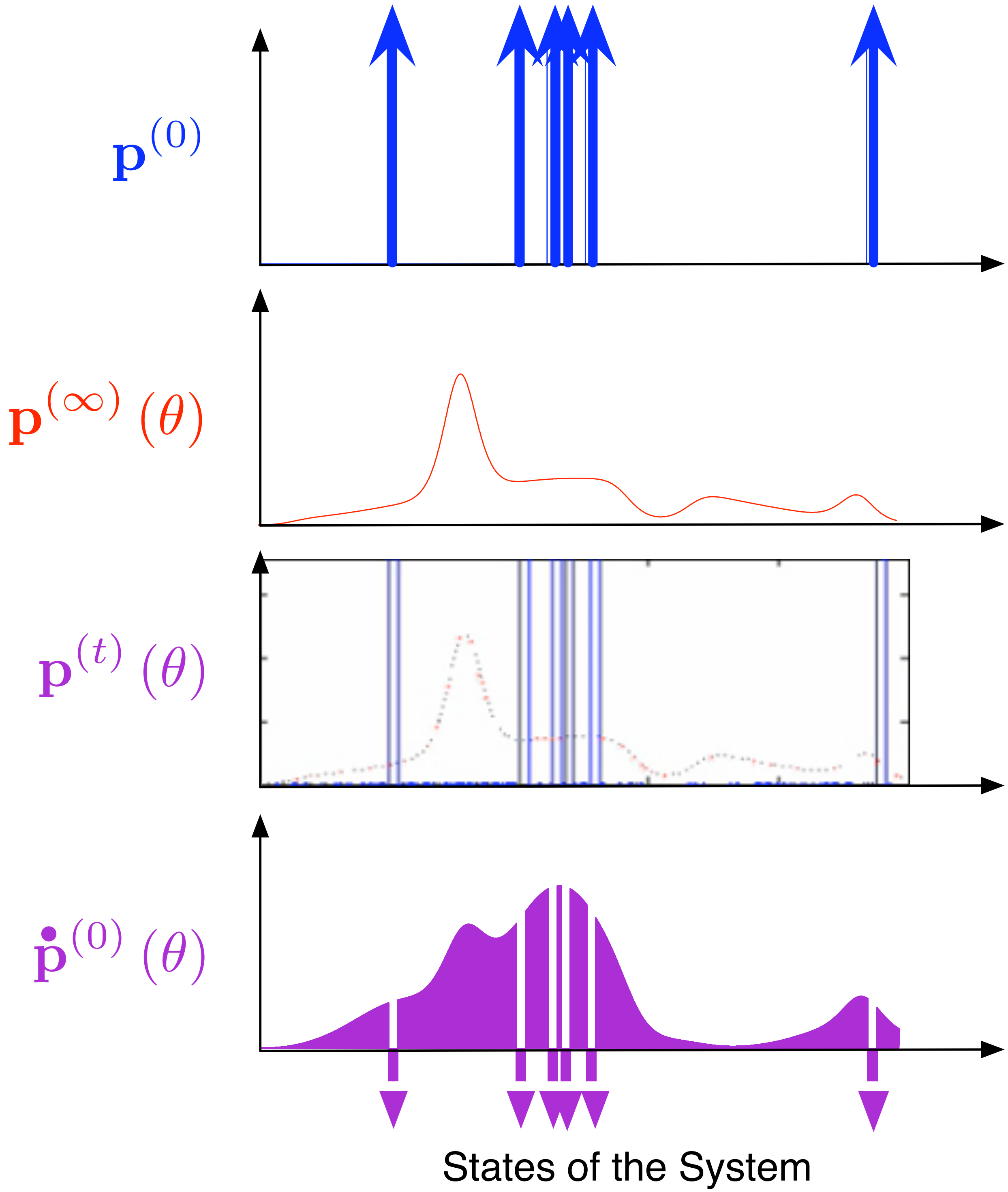
$$D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\infty)}(\theta)\right) \leq D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\epsilon)}(\theta)\right) + D_{KL}\left(\mathbf{p}^{(\epsilon)}(\theta)||\mathbf{p}^{(\infty)}(\theta)\right)$$

$$D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\infty)}(\theta)\right) - D_{KL}\left(\mathbf{p}^{(\epsilon)}(\theta)||\mathbf{p}^{(\infty)}(\theta)\right) \leq D_{KL}\left(\mathbf{p}^{(\mathbf{0})}||\mathbf{p}^{(\epsilon)}(\theta)\right)$$

$$K_{CD} \leq K_{MPF}$$

Alternative view

- Dynamics turn data distribution into model distribution
- Objective is to minimize initial flow of probability away from data - the shaded area

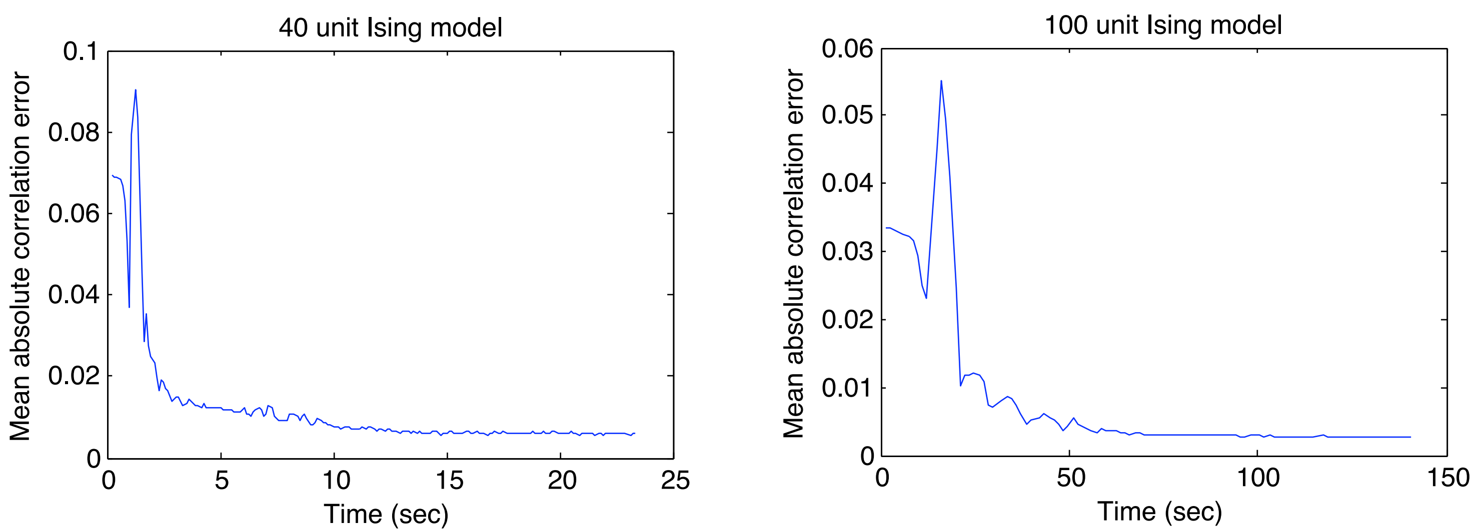


Examples - Ising

T Broderick, M Dudík, G Tkačik, R Schapire, and W Bialek. Faster solutions of the inverse pairwise ising problem. *E-print arXiv*, Jan 2007.

- Takes Broderick et al ~200 seconds on ~100 cores to recover parameters for 40 unit Ising model from 20,000 samples

- Using their J matrix, takes MPF ~15 seconds on 8 cores

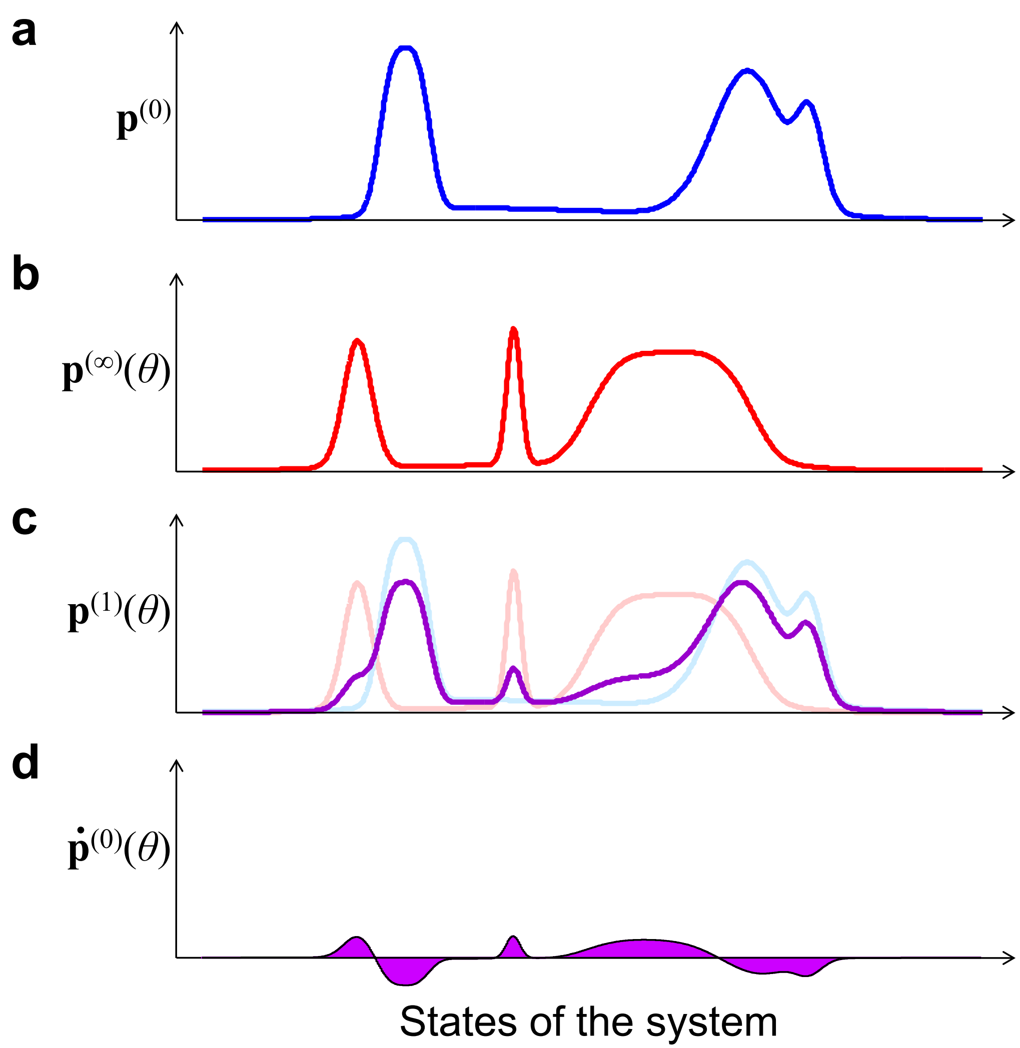


- Learning is ~ 2 orders of magnitude faster

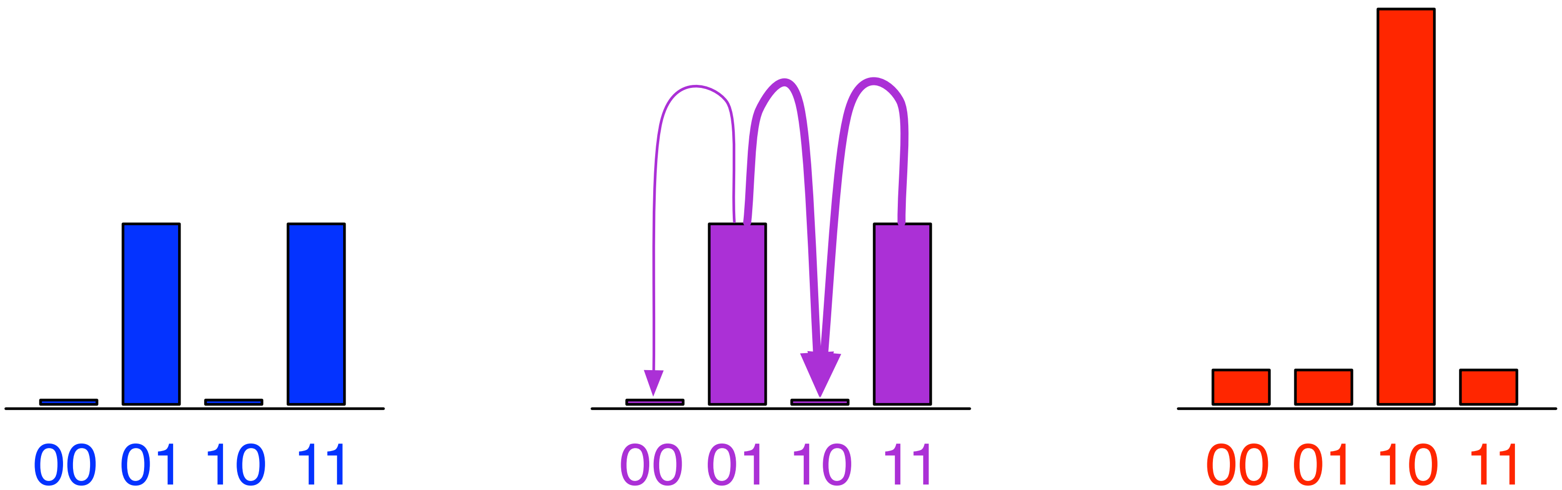
Thank you!

Objective function
Alternate interpretation

- Dynamics turn data distribution *(a)* into model distribution *(b)*
- *(c)* shows distribution at intermediate time
- The objective is to minimize the initial flow of probability away from the data, the shaded area in *(d)*.



MPF - Dynamics



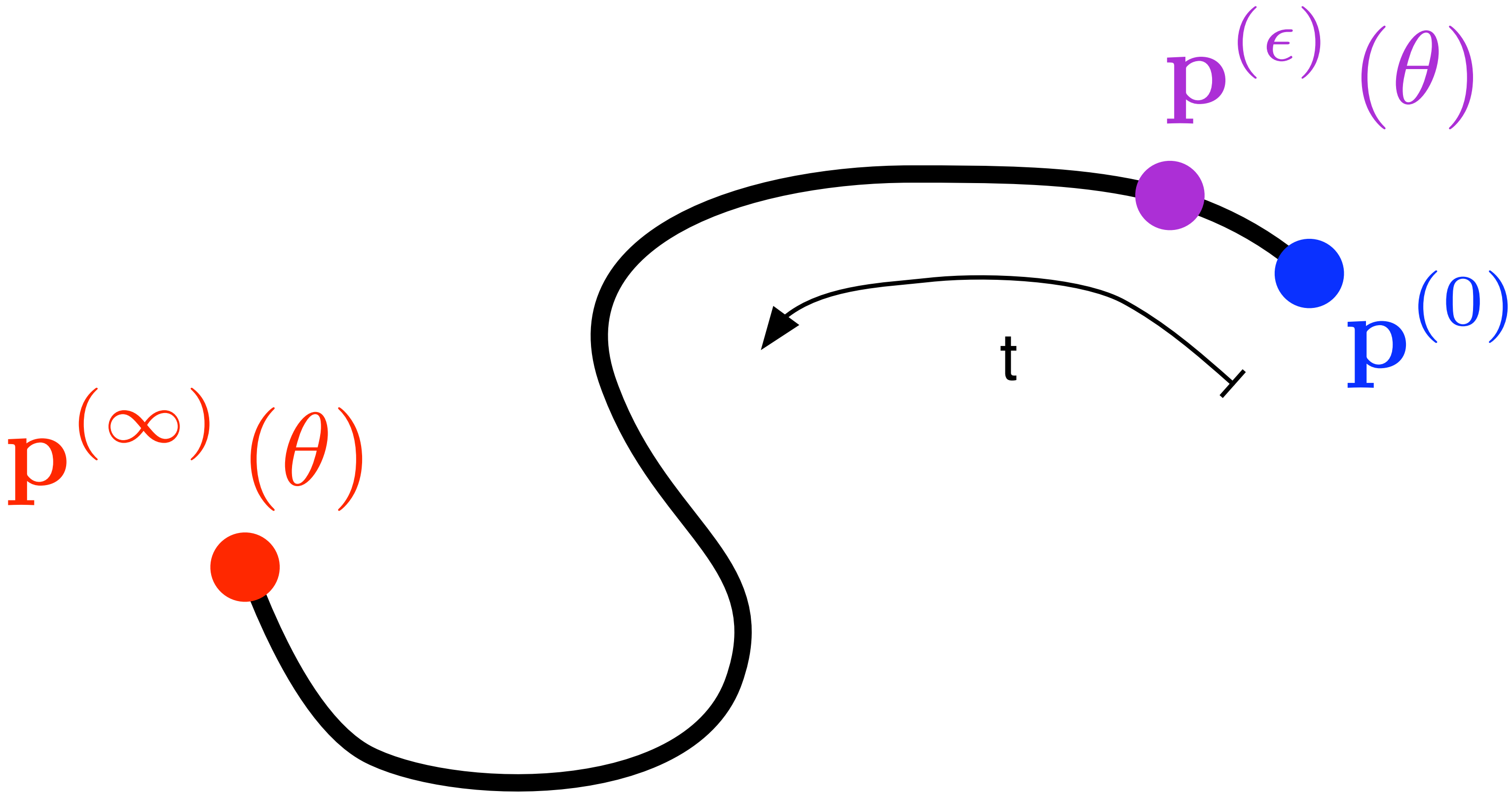
data distribution | dynamics | model distribution

$$p_i^{(0)} = \text{data} \qquad\qquad p_i^{(\infty)}(\theta) = \frac{e^{-E_i(\theta)}}{Z(\theta)}$$

$$\dot{p}_i^{(0)} = \sum_j \Gamma_{ij}(\theta)\, p_j^{(0)} \qquad \dot{p}_i^{(t)} = \sum_j \Gamma_{ij}(\theta)\, p_j^{(t)}(\theta) \qquad \dot{p}_i^{(\infty)} = 0$$

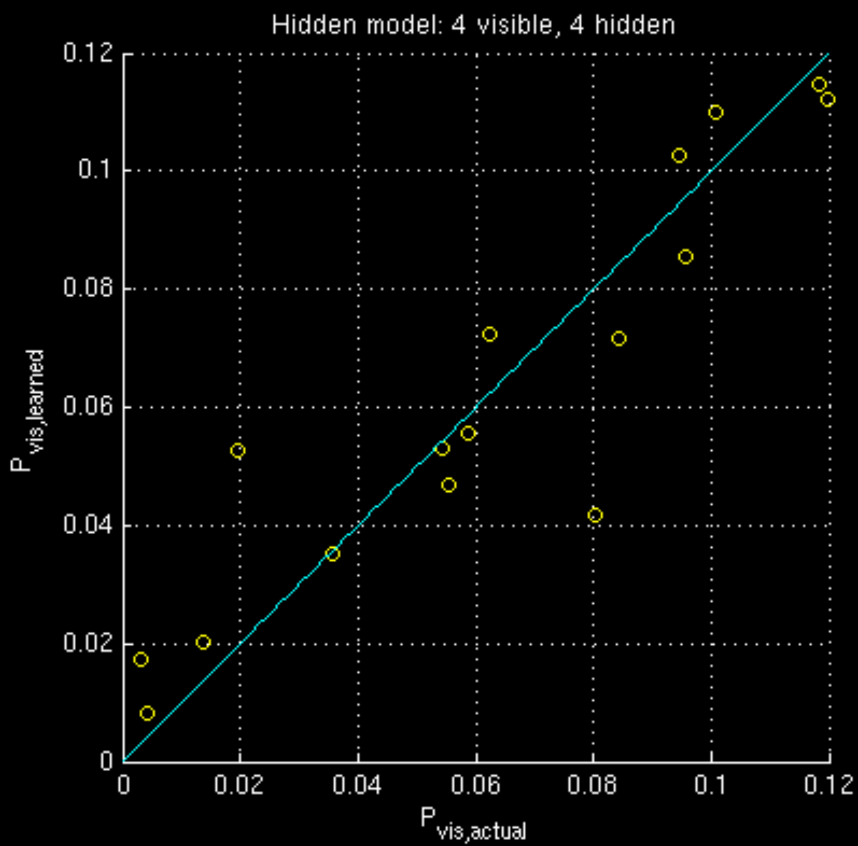
- Most Monte Carlo methods implement a stochastic version of these dynamics

Minimum probability flow
Overview

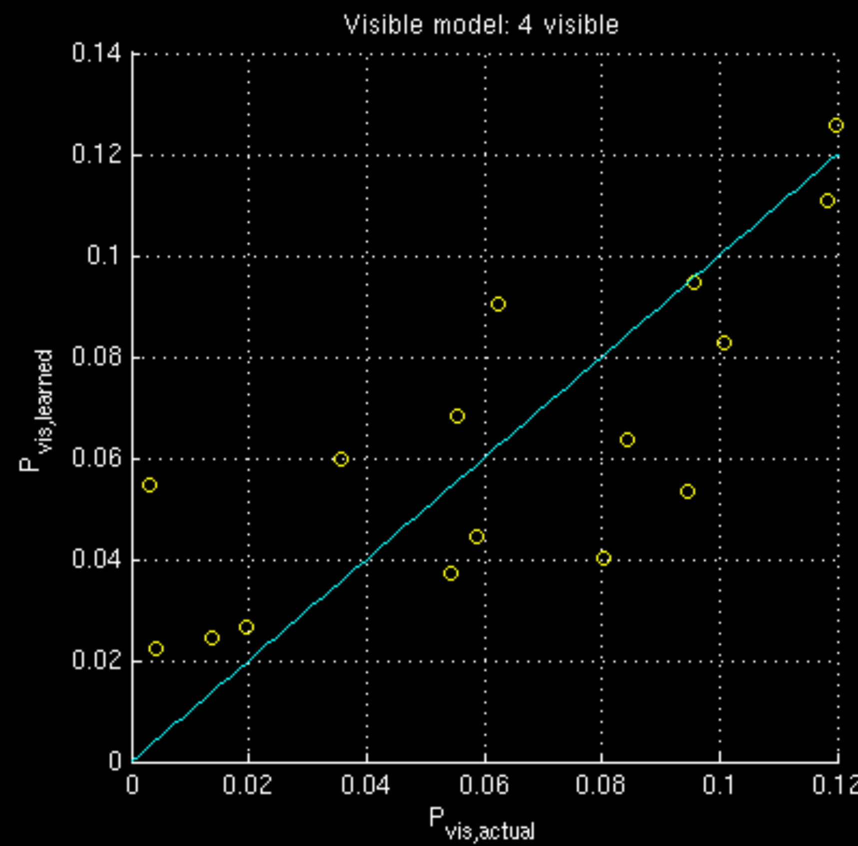


Example: Boltzmann Machine

Comparison of actual visible state probabilities:
4 visible, 4 hidden VS. only 4 visible



Hidden model



Visible model