Latent variables and Restricted Boltzmann Machines

Latent variables

enhance expressive power of generative models by encoding complex correlations between data

K-means

$r_{nm} \in 0, 1 \quad \leftarrow z$

$\mu_m \quad \leftarrow \theta$

Latent variables

enhance expressive power of generative models by encoding complex correlations between data

$z \longrightarrow h$ for "hidden" in this case

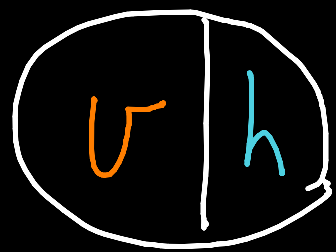
$x \longrightarrow v$ for "visible"

Latent variables

enhance expressive power of generative models by encoding complex correlations between data

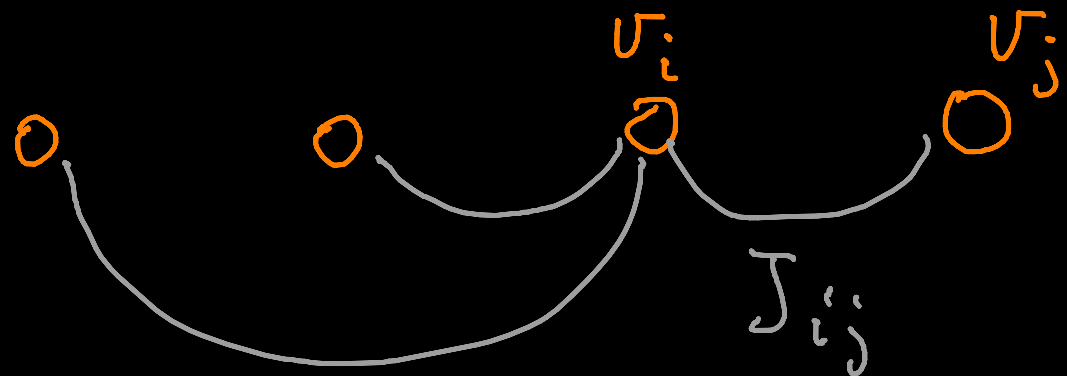
$z \to \underline{h}$ for "hidden" in this case

$x \to \underline{v}$ for "visible"



$v \cup h$ system

- spin systems (physics again relevant for UL ...)

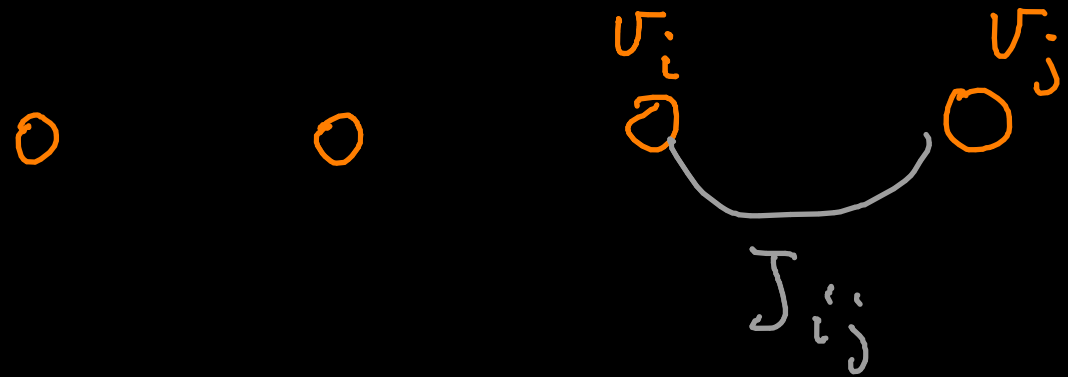


i = index of the spin (j ..)

mean field: all couplings $J_{ij} \neq 0$

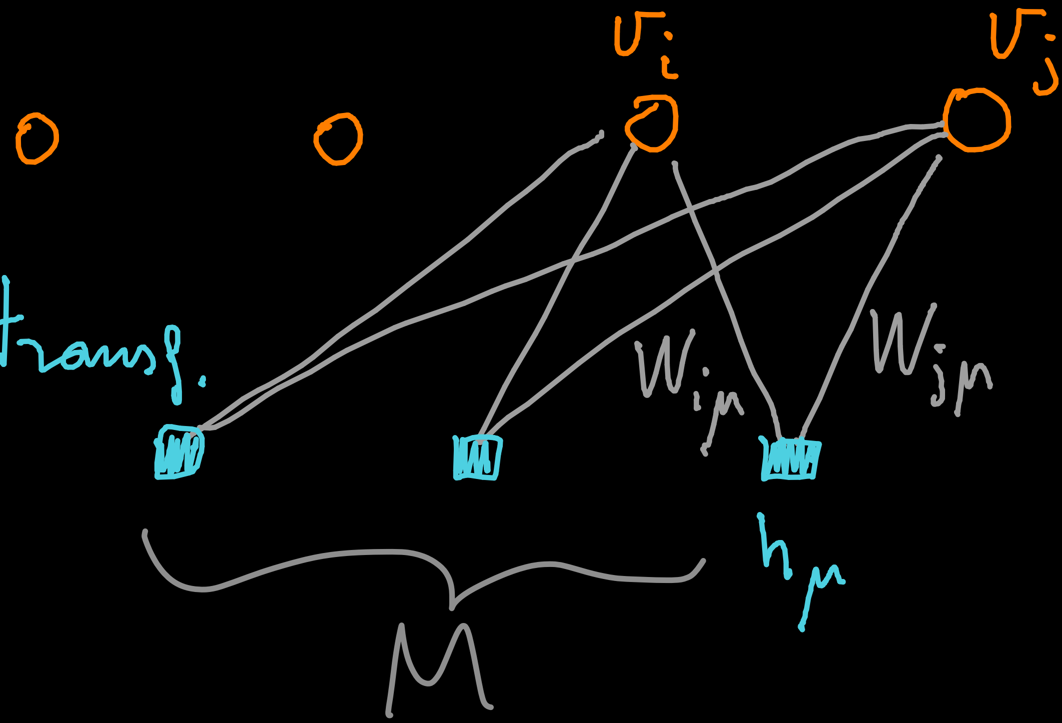
energy $$E(\sigma) = -\sum_i a_i \sigma_i - \frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j$$

$$J_{ij} = \sum_{\mu=1}^{M} W_{i\mu} W_{\mu j}$$



$$E(v) = -\sum_i a_i v_i - \frac{1}{2} \sum_{ij} J_{ij} v_i v_j$$

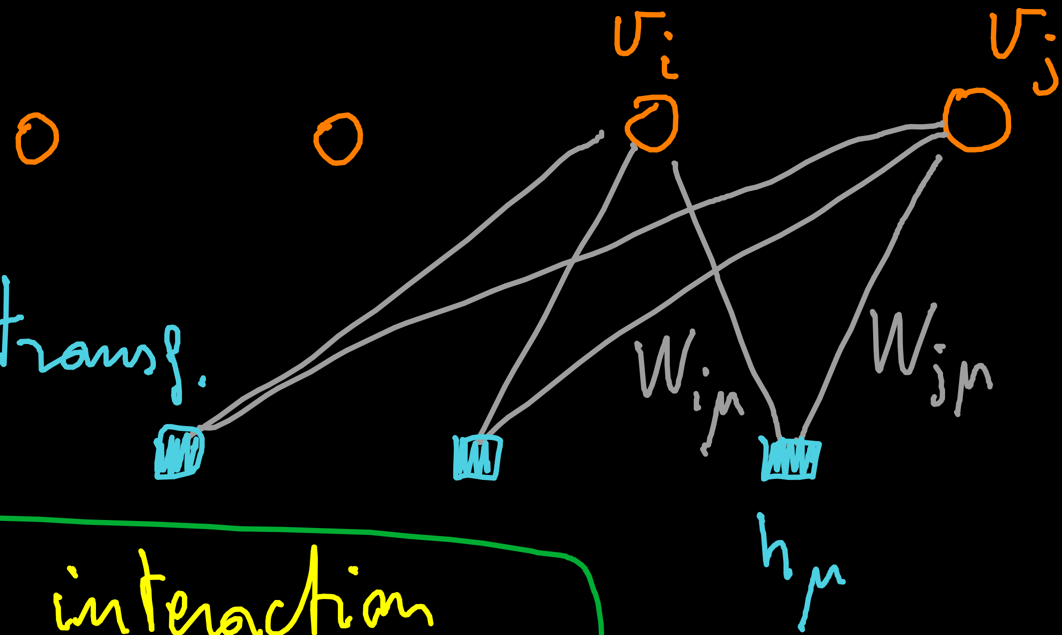
$$J_{ij} = \sum_{\mu=1}^{M} W_{i\mu} W_{j\mu}$$



Hubbard-Stratonovich transf.
(h_μ's with Gaussian stat.)

$$E(v) = -\sum_i a_i v_i - \frac{1}{2}\sum_{ij} J_{ij} v_i v_j$$

$$J_{ij} = \sum_\mu W_{i\mu} W_{j\mu}$$



Hubbard-Stratonovich transf.

$\boxed{J_{ij} \text{ removed: no direct interaction between "spins" } v_i \text{ \& } v_j}$

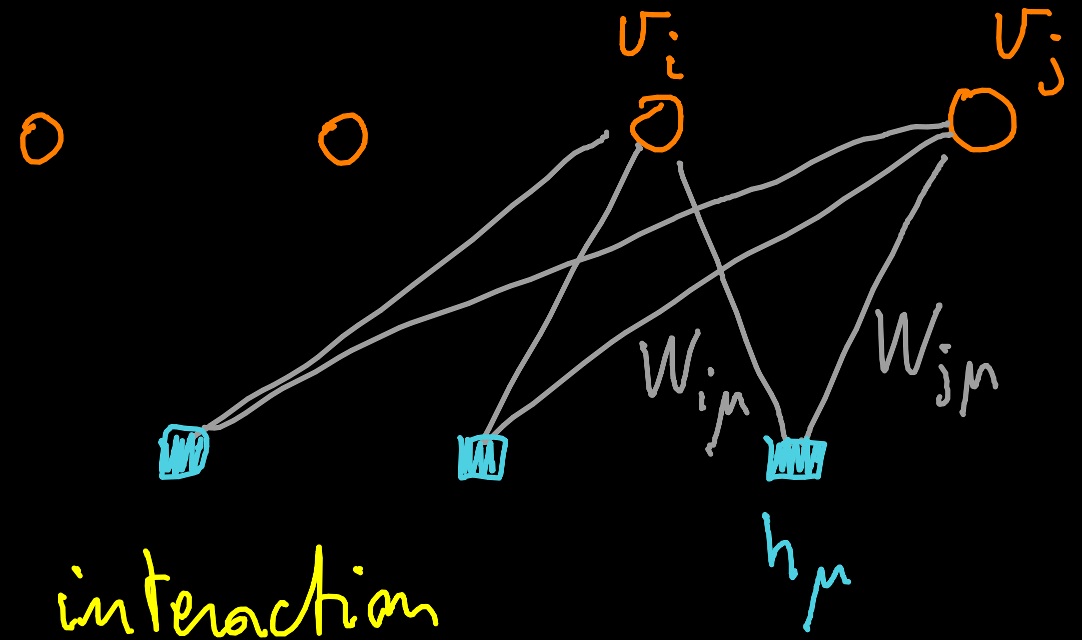
$$E(v) = -\sum_i a_i v_i - \frac{1}{2}\sum_{ij} J_{ij} v_i v_j$$

$$\Downarrow$$

$$E(v,h) = -\sum_i a_i v_i + \frac{1}{2}\sum_\mu h_\mu^2 - \sum_{i\mu} v_i W_{i\mu} h_\mu$$

visible layer

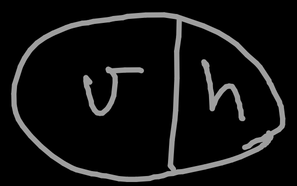
hidden layer



J_{ij} removed : no direct interaction between "spins" v_i & v_j

"Restricted"

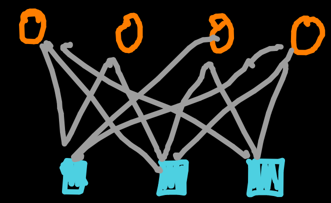
(also NO $h_\mu h_\nu$ interaction)



bipartite system

$$E(v,h) = -\sum_i a_i v_i + \frac{1}{2}\sum_\mu h_\mu^2 - \sum_{i\mu} v_i W_{i\mu} h_\mu$$

Restricted Boltzmann Machines



inspired by previous considerations,

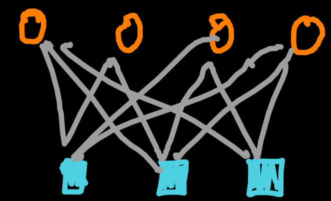
energy
$$\boxed{E(v,h) = -\sum_i a_i(v_i) - \sum_\mu b_\mu(h_\mu) - \sum_{i\mu} v_i W_{i\mu} h_\mu}$$

functions

$a_i(\cdot)$

$b_\mu(\cdot)$

Restricted Boltzmann Machines



inspired by previous considerations,

energy $\boxed{E(v,h) = -\sum_i a_i(v_i) - \sum_\mu b_\mu(h_\mu) - \sum_{i\mu} v_i W_{i\mu} h_\mu}$

functions

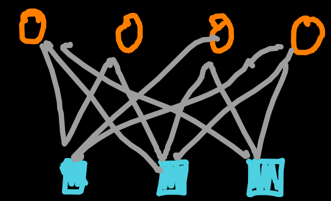
$a_i(\cdot)$

$b_\mu(\cdot)$

	Bernoulli Layers binary $v_i \in \{0,1\}$	Gaussian $v_i \in \mathbb{R}$
$a_i(v_i)$	$a_i v_i$	$\frac{v_i^2}{2\sigma_i^2}$
$b_\mu(h_\mu)$	$b_\mu h_\mu$	$\frac{h_\mu^2}{2\sigma_\mu^2}$

(also other versions, see Monasson et al.)

Restricted Boltzmann Machines



inspired by previous considerations,

energy $$\boxed{E(v,h) = -\sum_i a_i(v_i) - \sum_\mu b_\mu(h_\mu) - \sum_{i\mu} v_i W_{i\mu} h_\mu}$$

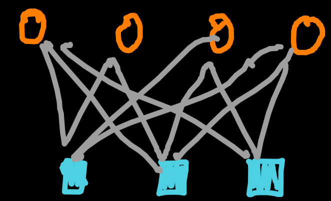
functions

$a_i(\cdot)$

$b_\mu(\cdot)$

	Bernoulli Layers binary $v_i \in \{0,1\}$	Gaussian $v_i \in \mathbb{R}$
$a_i(v_i)$	$a_i v_i$	$\frac{v_i^2}{2\sigma_i^2}$
$b_\mu(h_\mu)$	$b_\mu h_\mu$	$\frac{h_\mu^2}{2\sigma_\mu^2}$

Restricted Boltzmann Machines



energy $$\boxed{E(v,h) = -\sum_i a_i v_i - \sum_\mu b_\mu h_\mu - \sum_i v_i W_{i\mu} h_\mu}$$

$$v_i\,,\ h_\mu = -1, 1$$

OR

$$= 0, 1$$

Correlations induced by latent variables $\longrightarrow$ see the review

training

parameters $\theta = \{ W_{i\mu}, \; a_i, \; b_\mu \}$

$$O_j = \partial_{\theta_j} E_\theta(v,h) \qquad O_j(x) = O_j(v,h)$$

$$\partial_{\theta_j}\left(-\mathcal{L}(\theta)\right) = \langle O_j \rangle_{data} - \langle O_j \rangle_{model} \quad (195)$$

for example $\partial_{W_{i\mu}} E = -v_i h_\mu$ thanks to the simple linear appearance of term $v_i W_{i\mu} h_\mu$

hence training via (195) follows these gradient components of $-\mathcal{L}(\theta)$ to minimize it:

$$-\partial_{W_{i\mu}} \mathcal{L} = \langle -v_i h_\mu \rangle_{\text{data}} - \langle -v_i h_\mu \rangle_{\text{model}}$$

$$-\partial_{a_i} \mathcal{L} = \langle -v_i \rangle_{\text{data}} - \langle -v_i \rangle_{\text{model}}$$

$$-\partial_{b_\mu} \mathcal{L} = \langle -h_\mu \rangle_{\text{data}} - \langle -h_\mu \rangle_{\text{model}}$$

hence training via (195) follows these gradient components of $\mathcal{L}(\theta)$ to maximize

$$\partial_{W_{i\mu}} \mathcal{L} = \langle v_i h_\mu \rangle_{data} - \langle v_i h_\mu \rangle_{model}$$

$$\partial_{a_i} \mathcal{L} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model}$$

$$\partial_{b_\mu} \mathcal{L} = \langle h_\mu \rangle_{data} - \langle h_\mu \rangle_{model}$$

maximize log-likelihood

$$\partial_{W_{i\mu}} \mathcal{L} = \langle v_i h_\mu \rangle_{data} - \langle v_i h_\mu \rangle_{model}$$

$$\partial_{a_i} \mathcal{L} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model}$$

$$\partial_{b_\mu} \mathcal{L} = \underbrace{\langle h_\mu \rangle_{data} - \langle h_\mu \rangle_{model}}$$

same interpretation: optimum where predictions of model match the averages from data

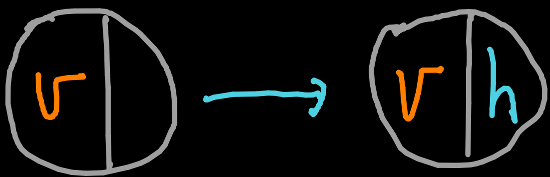
maximize log-likelihood

$$\partial_{W_{i\mu}} \mathcal{L} = \langle v_i h_\mu \rangle_{data} - \langle v_i h_\mu \rangle_{model}$$

$$\partial_{a_i} \mathcal{L} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model}$$

$$\partial_{b_\mu} \mathcal{L} = \underbrace{\langle h_\mu \rangle_{data}}_{\text{from "data"}} - \langle h_\mu \rangle_{model}$$

$v \cup h$



maximize log-likelihood

$$\partial_{W_{i\mu}} \mathcal{L} = \langle v_i h_\mu \rangle_{data} - \langle v_i h_\mu \rangle_{model}$$

$$\partial_{a_i} \mathcal{L} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model}$$

$$\partial_{b_\mu} \mathcal{L} = \langle h_\mu \rangle_{data} - \underbrace{\langle h_\mu \rangle_{model}}_{\text{run MC to generate } v' \,\&\, h'}$$

Gibbs sampling

much simplified by bipartite structure of restricted B.M. (no interaction between v's and between h's)

$\Rightarrow$ conditionally independent variables

Gibbs sampling

much simplified by bipartite structure of restricted B.M. (no interaction between v's and between h's)

$\Longrightarrow$ conditionally independent variables

$$p(v|h) = \prod_i p(v_i|h)$$

$$p(h|v) = \prod_\mu p(h_\mu|v)$$

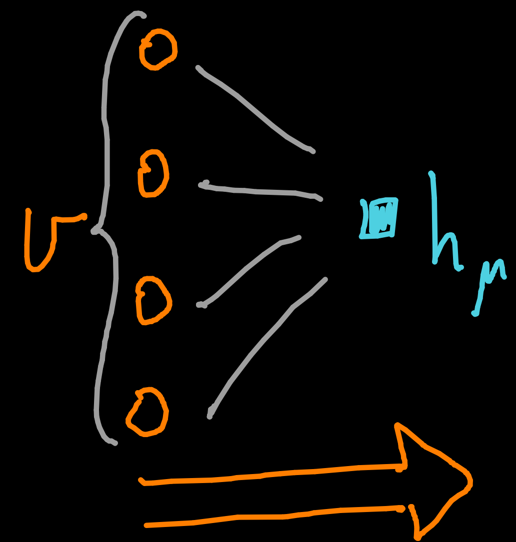
probabilities are factorized

$$\boxed{\begin{aligned} v &= \{v_i\} \\ h &= \{h_\mu\} \end{aligned}}$$

(212)

we can draw each h_μ independently from the others ("restricted"!) according to its

$$p(h_\mu | v)$$

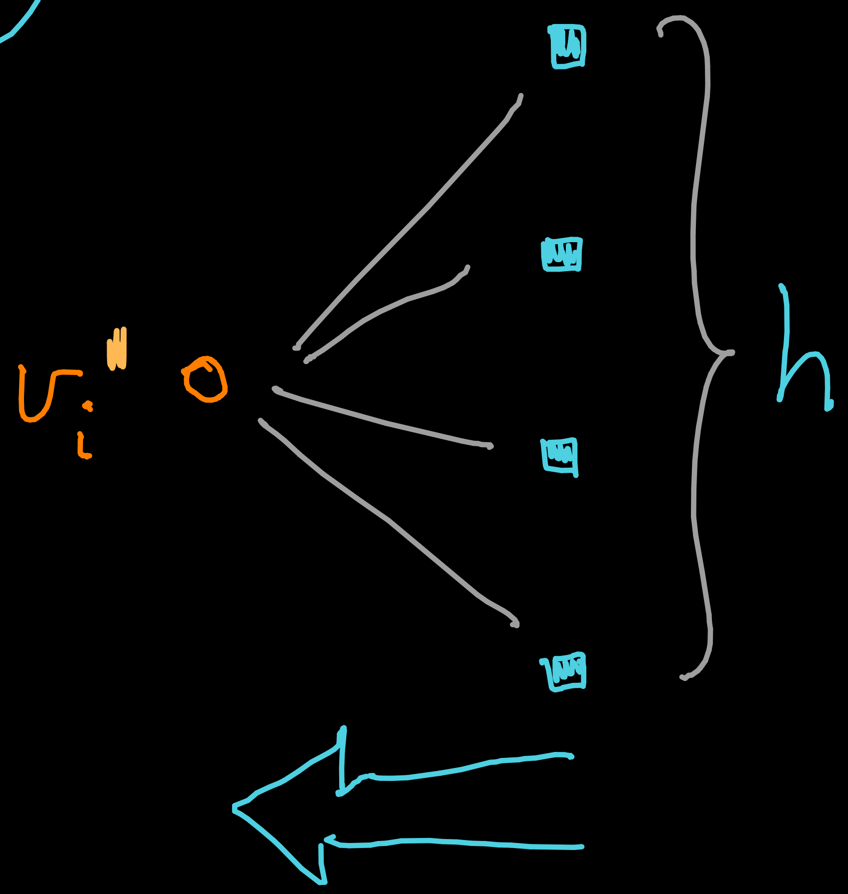


$$p(h|v) = \prod_\mu p(h_\mu | v)$$

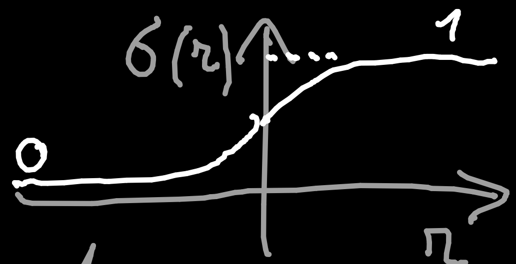
probabilities are factorized

We can draw each v_i' independently from the others ("restricted"!) according to its

$$p(v_i' \mid h)$$



for Bernoulli layers



(re-)defining sigmoid $\sigma(x) = \dfrac{1}{1+e^{-x}}$

if $v_i = 0, 1$

$$p(v_i = 1 \mid h) = \sigma\Big(a_i + \sum_\mu W_{i\mu} h_\mu\Big)$$

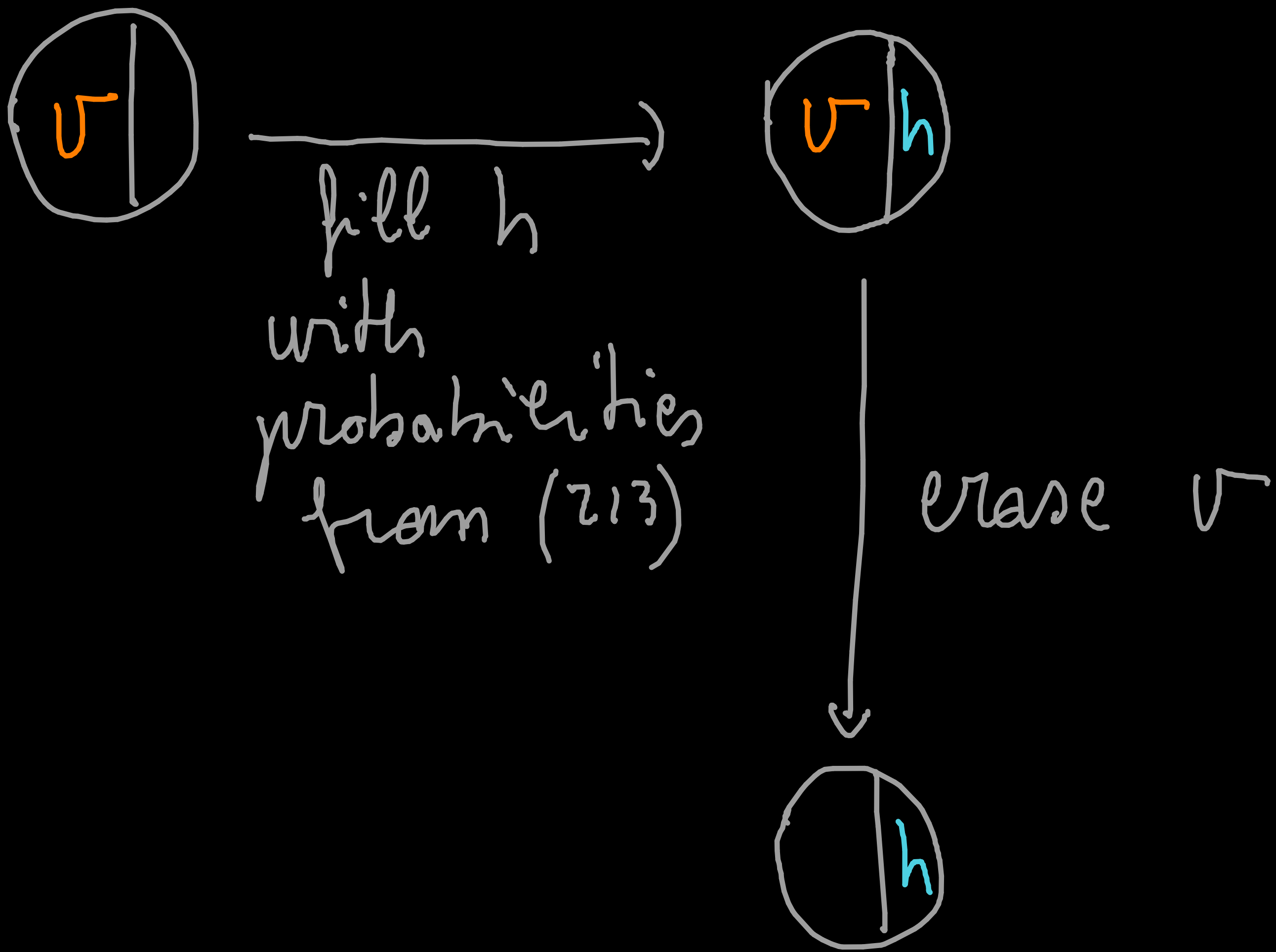
$$p(h_\mu = 1 \mid v) = \sigma\Big(b_\mu + \sum_i W_{i\mu} v_i\Big) \tag{213}$$

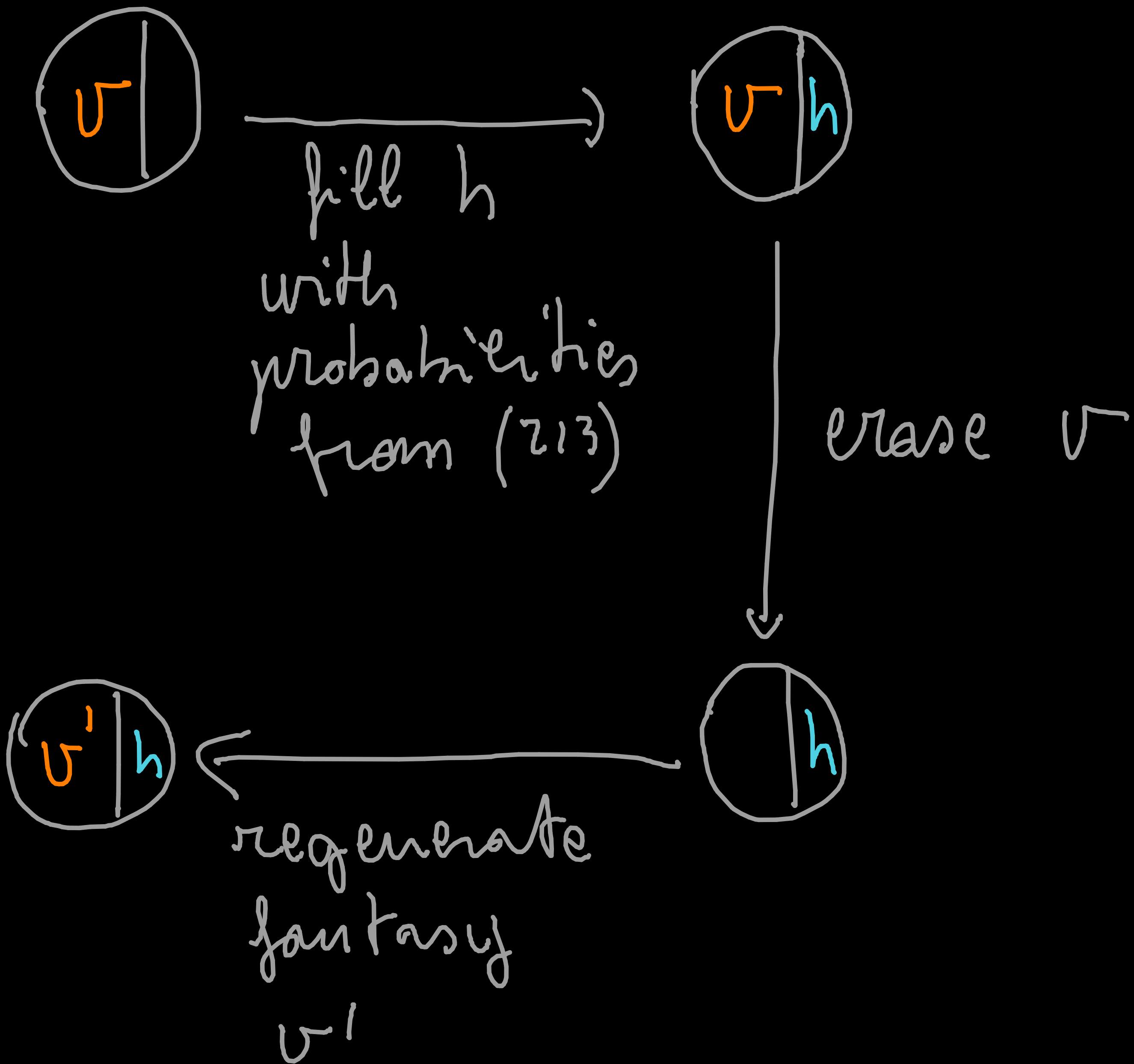
$v \,|\; \xrightarrow[\text{fill } h \text{ with probabilities from (213)}]{}$

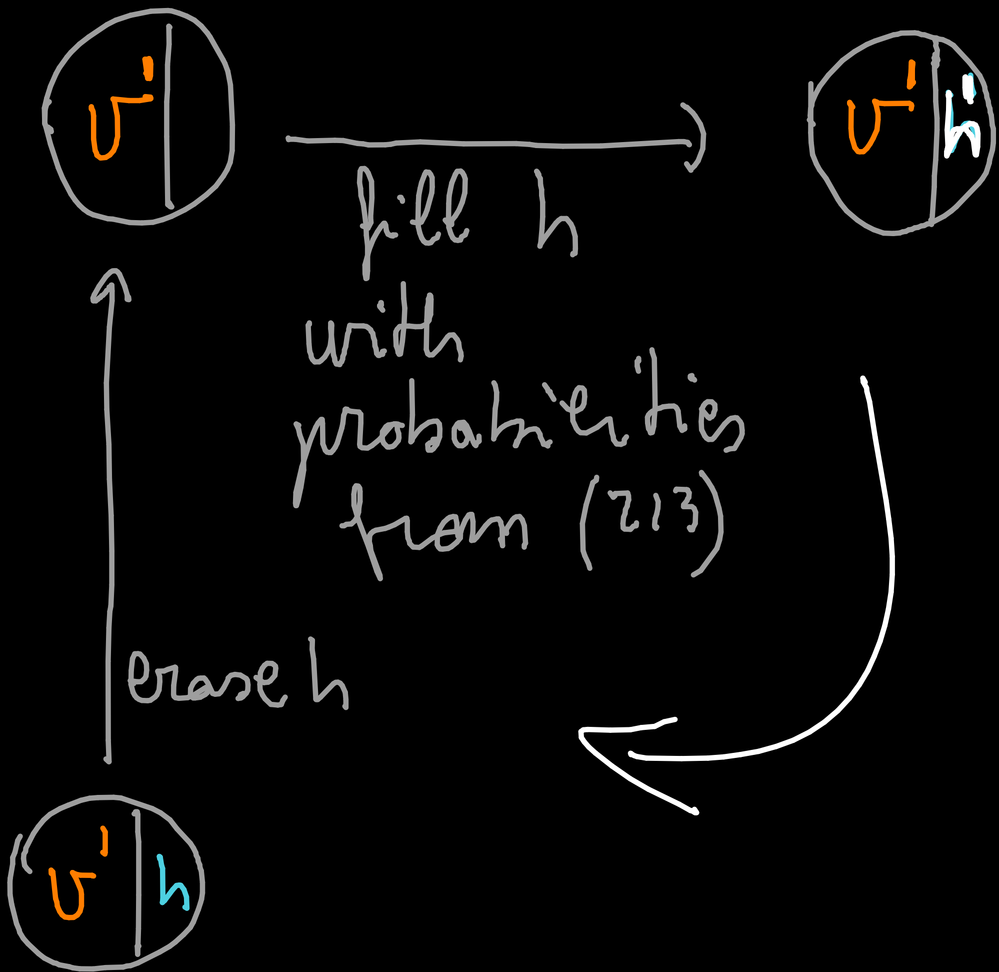
$$v \,|\, \quad \xrightarrow[\text{with probabilities from (213)}]{} \quad v \,|\, h$$

fill h

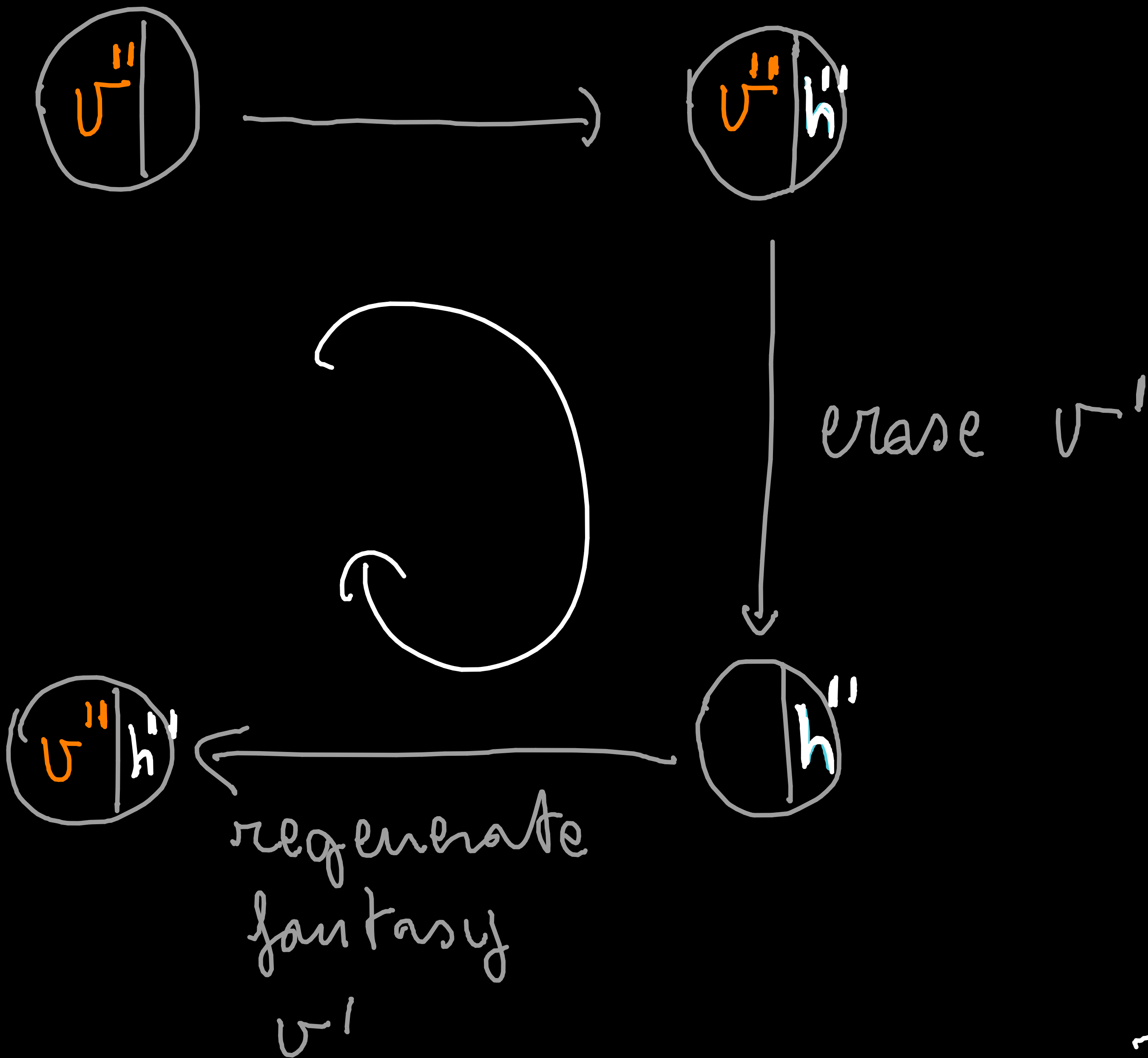
with probabilities from (213)


v
v h
fill h
with
probabilities
from (213)
erase v
h


v
v h
fill h with probabilities from (213)
erase v
h
regenerate fantasy v'
v' h



v'
v' | h'
fill h with probabilities from (213)
v' | h
erase h



repeat the process to sample the model averages

Alternating Gibbs sampling

$$
\begin{tikzcd}
v(0) \arrow[dr] & & v(1) \arrow[dr] & & v(2) \arrow[dr] & & \cdots \\
& h(0) \arrow[ur] & & h(1) \arrow[ur] & & h(2) \arrow[ur] &
\end{tikzcd}
$$

Alternating Gibbs sampling

$$
\begin{tikzcd}
v(0) \arrow[dr] & & v(1) \arrow[dr] & & v(2) \arrow[dr] & & \cdots \\
& h(0) \arrow[ur] & & h(1) \arrow[ur] & & h(2) \arrow[ur] &
\end{tikzcd}
$$

positive phase

negative phase

Alternating Gibbs sampling

$$v(0) \to h(0) \to v(1) \to h(1) \to v(2) \to h(2) \to \cdots$$

"data"

at $t = 0$

$$\downarrow$$

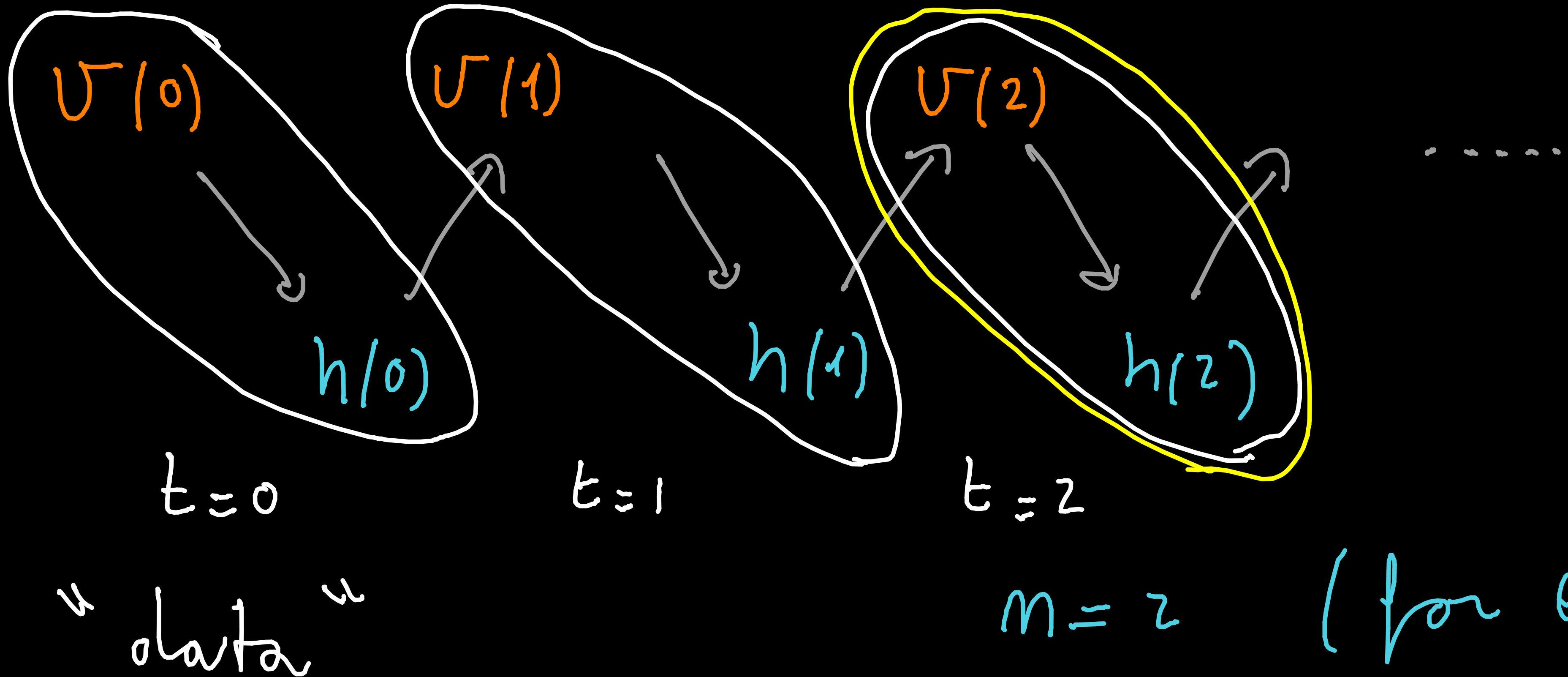
$$\langle \ldots\ldots \rangle_{data}$$

Alternating Gibbs sampling

$$v(0) \to h(0) \to v(1) \to h(1) \to v(2) \to h(2) \to \cdots\cdots$$

"model"
at $t \to \infty$

Contrastive Divergence (CD-n)

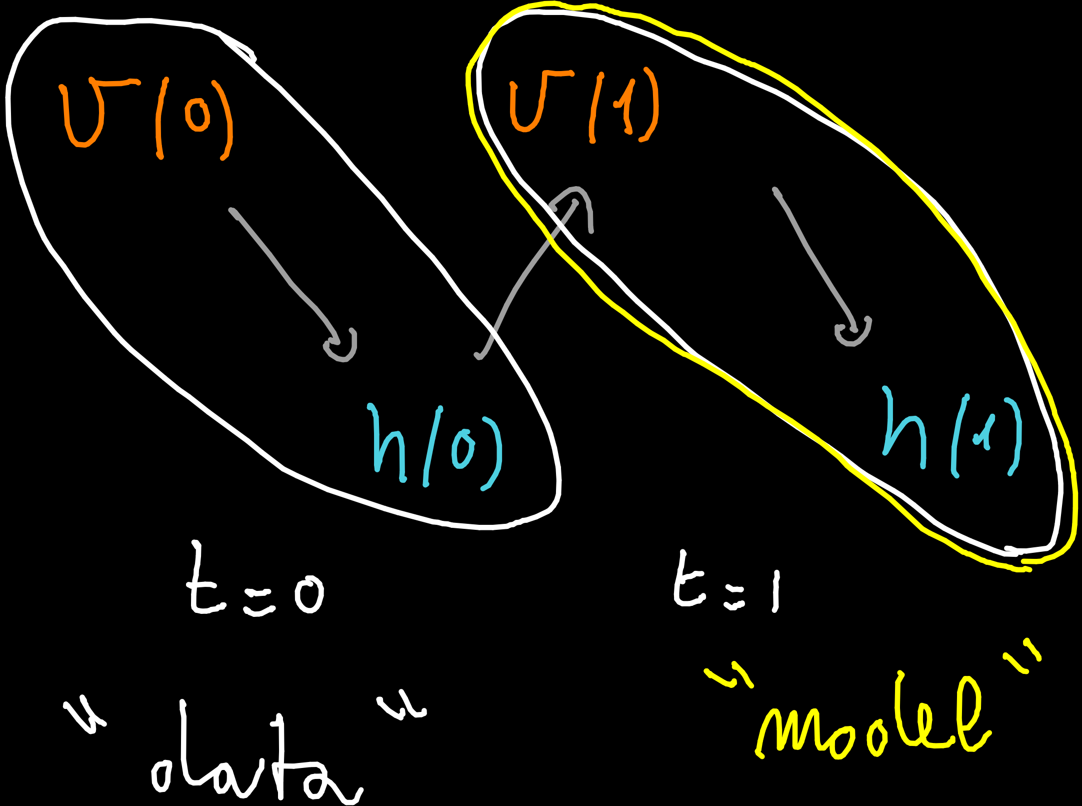


$n = 2$ (for example)

↓

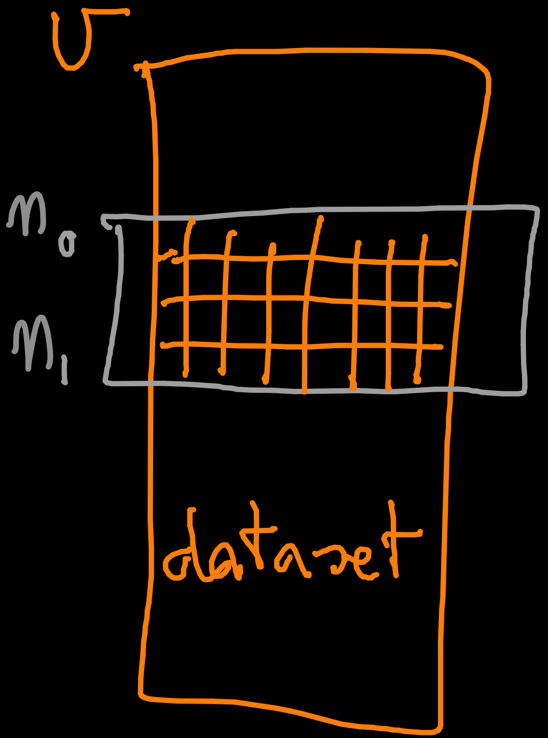
"model" evaluated at $t = n$ rather than $t \to \infty$

Contrastive Divergence (CD-1)



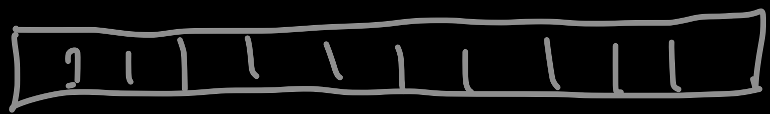
- most extreme example of CD
- fastest
- it works ...

Mini batches

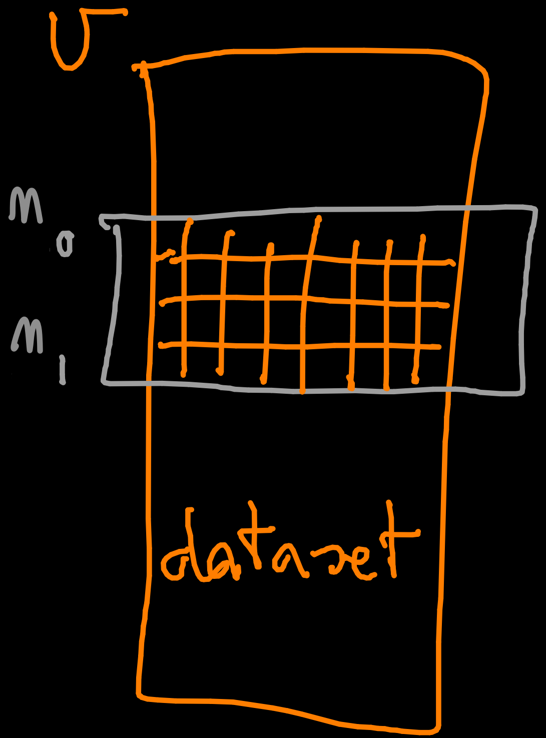


$$B = \{ v^{(n_0)}, \dots, v^{(n_1)} \}$$

$$v^{(n)} = \left(v_{i=1}, v_{i=2}, \dots v_{i=L} \right)^{(n)}$$

n 

Mini batches



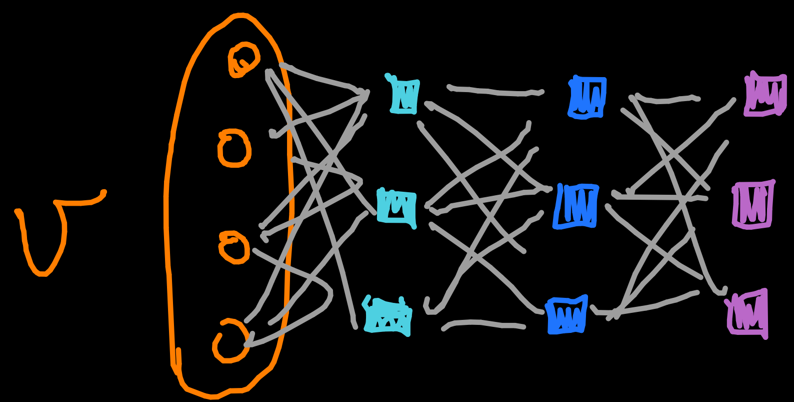
$$B = \{ v^{(n_0)}, \dots, v^{(n_1)} \}$$

$$v^{(n)} = \left(v_{i=1}, v_{i=2}, \dots v_{i=L} \right)^{(n)}$$

$$\left.\begin{array}{l} \langle \dots \rangle_{\text{data}} \longrightarrow \langle \dots \rangle^{B}_{\text{data}} \\ \\ \langle \dots \rangle_{\text{model}} \longrightarrow \langle \dots \rangle^{B}_{\text{model}} \end{array}\right\} \text{gradient} \Longrightarrow \text{SGD method}$$

More reading in the review:

- initialization
- regularization
- learning rates
- persistent contrastive divergence
- deep Boltzmann machines
  (many hidden layers)



Summary: after training

- RBM has hidden layer that responds to data and can send back "fantasy data" with similar features
- <u>generative</u>
- denoising
- ...
- "reading" W's ⇒ understand data