Attention via $\log \sum \exp$ energy

Alexander Tschantz

January 30, 2025

Overview

$\log \sum \exp$ framework

$\log \sum \exp$ examples

$\log \sum \exp$ graphical models

$\log \sum \exp$ spatio-temporal model

Renormalised mixture models

$\log \sum \exp$ energy

We consider energy functions that map a set of parents $\{\boldsymbol{v}_p \in \mathbb{R}^{d_p} : p \in P\}$ to a set of children $\{\boldsymbol{v}_c \in \mathbb{R}^{d_c} : c \in C\}$.

Each energy function defines a similarity function (with parameters θ) that measures agreement between a parent and child vector:

$$\mathrm{sim}_\theta(\boldsymbol{v}_c, \boldsymbol{v}_p) : \mathbb{R}^{d_c} \times \mathbb{R}^{d_p} \to \mathbb{R}.$$

The $\log \sum \exp$ energy function is then given by:

$$E(\{\boldsymbol{v}_c\}, \{\boldsymbol{v}_p\}, \theta) = -\sum_{c \in C} \ln\Big(\sum_{p \in P} \exp\big(\mathrm{sim}_\theta(\boldsymbol{v}_c, \boldsymbol{v}_p)\big)\Big).$$

$\log \sum \exp$ derivatives

We define *attention* as:

$$\alpha_{c,p} = \mathsf{softmax}_p\big(\mathrm{sim}_\theta(\boldsymbol{v}_c, \boldsymbol{v}_p)\big) = \frac{\exp\big(\mathrm{sim}_\theta(\boldsymbol{v}_c, \boldsymbol{v}_p)\big)}{\sum_{p' \in P} \exp\big(\mathrm{sim}_\theta(\boldsymbol{v}_c, \boldsymbol{v}_{p'})\big)}.$$

The derivatives with are then given by:

$$-\frac{\partial E}{\partial \boldsymbol{v}_c} = \sum_{p \in P} \alpha_{c,p} \frac{\partial \mathrm{sim}_\theta(\boldsymbol{v}_c, \boldsymbol{v}_p)}{\partial \boldsymbol{v}_c},$$

$$-\frac{\partial E}{\partial \boldsymbol{v}_p} = \sum_{c \in C} \alpha_{c,p} \frac{\partial \mathrm{sim}_\theta(\boldsymbol{v}_c, \boldsymbol{v}_p)}{\partial \boldsymbol{v}_p},$$

$$-\frac{\partial E}{\partial \theta} = \sum_{c \in C} \sum_{p \in P} \alpha_{c,p} \frac{\partial \mathrm{sim}_\theta(\boldsymbol{v}_c, \boldsymbol{v}_p)}{\partial \theta}.$$

$\log \sum \exp$ derivatives

The gradient of the energy function can be interpreted as an *expected value* over a discrete distribution $\alpha_{c,p} = P(p \mid c)$:

$$-\frac{\partial E}{\partial \boldsymbol{v}_c} = \mathbb{E}_{p \sim P(p|c)} \left[\frac{\partial \mathrm{sim}_\theta(\boldsymbol{v}_c, \boldsymbol{v}_p)}{\partial \boldsymbol{v}_c} \right] \tag{1}$$

While the child gradient is an exact expectation under $P(p \mid c)$, the parent gradient does not correspond exactly to an expectation under $P(c \mid p)$; it is instead proportional to it through Bayes rule.

$\log \sum \exp$ graphical models

We consider a single set of N nodes $\boldsymbol{v} = \{\boldsymbol{v}_i : i \in \{1, 2, \ldots, N\}\}$, where each node $\boldsymbol{v}_i \in \mathbb{R}^{d_i}$, and M energy functions $\{E_m : m \in \{1, 2, \ldots, M\}\}$.

Each E_m has a similairtiy function $\mathrm{sim}_m(\cdot)$ with parameters θ_m and defines a subset of nodes acting as *children* $C_m \subseteq \{1, 2, \ldots, N\}$ and a subset acting as *parents* $P_m \subseteq \{1, 2, \ldots, N\}$, which may overlap.

$$E_m\big(\{\boldsymbol{v}_c\}, \{\boldsymbol{v}_p\}, \theta_m\big) = -\sum_{c \in C_m} \ln\Big(\sum_{p \in P_m} \exp\big(\mathrm{sim}_m(\boldsymbol{v}_c, \boldsymbol{v}_p)\big)\Big).$$

The full energy is the sum over all energy functions:

$$E(\{\boldsymbol{v}_i\}, \{\theta_m\}) = \sum_{m=1}^{M} E_m\big(\{\boldsymbol{v}_c\}, \{\boldsymbol{v}_p\}, \theta_m\big).$$

$\log \sum \exp$ expectation maximisation

Expectation Step:

- Perform gradient descent on each $\boldsymbol{v}_i$ to find the optimal latent states:

$$\boldsymbol{v}_i^* = \arg\min_{\boldsymbol{v}_i} E(\{\boldsymbol{v}_i\}, \{\theta_m\}).$$

Maximisation Step:

- Given the updated $\boldsymbol{v}_i^*$, take a gradient step on each θ_m:

$$\theta_m = \theta_m - \eta \frac{\partial E}{\partial \theta_m}\bigg|_{\boldsymbol{v}_i^*}.$$

log $\sum$ exp framework

```
class Energy:
    def similarity(self, *args):
        raise NotImplementedError

    def __call__(self, *args):
        # energy function
        sim_matrix = self.similarity(*args)
        return -sum(logsumexp(sim_matrix, axis=1))

dx = grad(energy)(x, y)
```

Gaussian Mixture Model

We define a set of child nodes $\{\boldsymbol{x}_i\}_{i=1}^N$ and parent means $\{\boldsymbol{\mu}_k\}_{k=1}^K$. The covariance matrices $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ are treated as parameters $\theta_k = \boldsymbol{\Sigma}_k$ (note that we could have also included means in the similarity parameters).

The similarity function is given by:

$$\mathrm{sim}_k\big(\boldsymbol{x}_i, \boldsymbol{\mu}_k\big) \;=\; -\tfrac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k).$$

The energy function is:

$$E^{\mathrm{GMM}} = -\sum_{i=1}^{N} \ln\Big(\sum_{k=1}^{K} \exp\big(\mathrm{sim}_k(\boldsymbol{x}_i, \boldsymbol{\mu}_k)\big)\Big).$$

The similarity function can be interpreted as the log probability of $\boldsymbol{x}_i$ under the conditional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, up to a normalization constant.

Gaussian Mixture Model

The gradient for $\boldsymbol{\mu}_k$ is given by:

$$-\frac{\partial \boldsymbol{E}^{\mathrm{GMM}}}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} \underbrace{\mathrm{softmax}_k\left(\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k)\right)}_{\alpha_{i,k}} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k).$$

Solving for zero gives the fixed point update for $\boldsymbol{\mu}_k$:

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \alpha_{i,k}\, \boldsymbol{x}_i}{\sum_{i=1}^{N} \alpha_{i,k}}.$$

Hopfield Attention

We define a set of child nodes $\{\boldsymbol{x}_i\}_{i=1}^N$, each $\boldsymbol{x}_i \in \mathbb{R}^d$, and a set of parent memory vectors $\{\boldsymbol{m}_\mu\}_{\mu=1}^K$, each $\boldsymbol{m}_\mu \in \mathbb{R}^d$.

The energy function is given by:

$$\mathrm{sim}\big(\boldsymbol{x}_i, \boldsymbol{m}_\mu\big) = \boldsymbol{x}_i^\top \boldsymbol{m}_\mu.$$

$$E^{\mathrm{Hopfield}} = -\sum_{i=1}^N \ln\Big(\sum_{\mu=1}^K \exp\big(\boldsymbol{x}_i^\top \boldsymbol{m}_\mu\big)\Big).$$

The gradients are:

$$\alpha_{i,\mu} = \mathrm{softmax}_\mu\big(\boldsymbol{x}_i^\top \boldsymbol{m}_\mu\big).$$

$$-\frac{\partial E}{\partial \boldsymbol{m}_\mu} = \sum_{i=1}^N \alpha_{i,\mu}\, \boldsymbol{x}_i, \qquad -\frac{\partial E}{\partial \boldsymbol{x}_i} = \sum_{\mu=1}^K \alpha_{i,\mu}\, \boldsymbol{m}_\mu,$$

Slot Attention

We define a set of child (token) nodes $\{\boldsymbol{x}_j\}_{j=1}^N$ and a set of parent (slot) nodes $\{\boldsymbol{\mu}_i\}_{i=1}^S$. The parameters θ consist of two projection matrices $\boldsymbol{W}_K$ and $\boldsymbol{W}_Q$.
The energy function is:

$$\mathrm{sim}\big(\boldsymbol{x}_j, \boldsymbol{\mu}_i\big) = (\boldsymbol{W}_K \boldsymbol{x}_j)^\top (\boldsymbol{W}_Q \boldsymbol{\mu}_i).$$

$$E^{\mathrm{Slot}} = -\sum_{j=1}^{N} \ln\Bigg(\sum_{i=1}^{S} \exp\big(\mathrm{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\big)\Bigg).$$

The gradients are:

$$\alpha_{j,i} = \mathrm{softmax}_i\big(\mathrm{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\big).$$

$$-\frac{\partial E}{\partial \boldsymbol{\mu}_i} = \sum_{j=1}^{N} \alpha_{j,i}\, \boldsymbol{W}_Q^\top \boldsymbol{W}_K\, \boldsymbol{x}_j, \qquad -\frac{\partial E}{\partial \boldsymbol{x}_j} = \sum_{i=1}^{S} \alpha_{j,i}\, \boldsymbol{W}_K^\top \boldsymbol{W}_Q\, \boldsymbol{\mu}_i.$$

Self-Attention

We define a set of nodes $\{\boldsymbol{x}_i\}_{i=1}^N$, where each $\boldsymbol{x}_i$ serves as both a child (query) and a parent (key). The parameters θ consist of projection matrices $\boldsymbol{W}^Q$ and $\boldsymbol{W}^K$, forming:

$$\boldsymbol{q}_i = \boldsymbol{W}^Q \boldsymbol{x}_i, \quad \boldsymbol{k}_i = \boldsymbol{W}^K \boldsymbol{x}_i.$$

The energy function is:

$$\mathrm{sim}\big(\boldsymbol{x}_c, \boldsymbol{x}_p\big) = \boldsymbol{q}_c^\top \boldsymbol{k}_p.$$

$$E^{\mathrm{SA}} = -\sum_{c=1}^{N} \ln\Big(\sum_{p=1}^{N} \exp(\boldsymbol{q}_c^\top \boldsymbol{k}_p)\Big).$$

The gradients are:

$$\alpha_{c,p} = \mathrm{softmax}_p\big(\boldsymbol{q}_c^\top \boldsymbol{k}_p\big).$$

$$-\frac{\partial E}{\partial \boldsymbol{x}_i} = \underbrace{\sum_{p=1}^{N} \alpha_{i,p} \boldsymbol{W}^{Q\top} \boldsymbol{W}^K \boldsymbol{x}_p}_{\text{Child side}} + \underbrace{\sum_{c=1}^{N} \alpha_{c,i} \boldsymbol{W}^{K\top} \boldsymbol{W}^Q \boldsymbol{x}_c}_{\text{Parent side}}.$$

Linear Mixture Model

We define a set of child nodes $\{\boldsymbol{x}_i\}_{i=1}^N$ and a set of parent nodes $\{\boldsymbol{z}_k\}_{k=1}^K$. The parameters $\boldsymbol{\theta} = \{(\boldsymbol{A}_k, \boldsymbol{b}_k, \Sigma_k)\}_{k=1}^K$ define a linear mapping:

$$\boldsymbol{x}_i = \boldsymbol{A}_k \boldsymbol{z}_k + \boldsymbol{b}_k + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma_k).$$

The energy function is:

$$\mathrm{sim}_k(\boldsymbol{x}_i, \boldsymbol{z}_k) = -\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{A}_k \boldsymbol{z}_k - \boldsymbol{b}_k)^\top \Sigma_k^{-1} (\boldsymbol{x}_i - \boldsymbol{A}_k \boldsymbol{z}_k - \boldsymbol{b}_k).$$

$$E^{\mathrm{LM}} = -\sum_{i=1}^N \ln\Big(\sum_{k=1}^K \exp\big(\mathrm{sim}_k(\boldsymbol{x}_i, \boldsymbol{z}_k)\big)\Big).$$

Note this is the energy function for the switching linear dynamical system (SLDS) model, and our simple latent attention (SLA) model.

Linear Mixture Model

The attention weights are:

$$\alpha_{i,k} = \mathrm{softmax}_k\big(\mathrm{sim}_k(\boldsymbol{x}_i, \boldsymbol{z}_k)\big).$$

The gradient w.r.t. $\boldsymbol{x}_i$ is:

$$-\frac{\partial E}{\partial \boldsymbol{x}_i} = \sum_{k=1}^{K} \alpha_{i,k} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{A}_k \boldsymbol{z}_k - \boldsymbol{b}_k).$$

The fixed-point update for $\boldsymbol{x}_i$ is:

$$\boldsymbol{x}_i^* = \sum_{k=1}^{K} \alpha_{i,k} \left(\boldsymbol{A}_k \boldsymbol{z}_k + \boldsymbol{b}_k\right).$$

Non-Linear Mixture Model

We define a set of child nodes $\{\boldsymbol{x}_i\}_{i=1}^{N}$ and a set of parent nodes $\{\boldsymbol{z}_k\}_{k=1}^{K}$. The parameters $\boldsymbol{\theta} = \{\boldsymbol{A}_k, \boldsymbol{b}_k, \sigma\}$ define a non-linear mapping:

$$f(\boldsymbol{z}_k) = \sigma(\boldsymbol{A}_k \boldsymbol{z}_k + \boldsymbol{b}_k),$$

where $\sigma(\cdot)$ is a non-linearity.

The similarity function is:

$$\mathrm{sim}_k(\boldsymbol{x}_i, \boldsymbol{z}_k) = -\tfrac{1}{2}\|\boldsymbol{x}_i - f(\boldsymbol{z}_k)\|^2.$$

The energy function is:

$$E^{\mathrm{NL}} = -\sum_{i=1}^{N} \ln\Big(\sum_{k=1}^{K} \exp\big(\mathrm{sim}_k(\boldsymbol{x}_i, \boldsymbol{z}_k)\big)\Big).$$

Non-Linear Mixture Model: Gradients

The attention weights are:

$$\alpha_{i,k} = \mathrm{softmax}_k\big(\mathrm{sim}_k(\boldsymbol{x}_i, \boldsymbol{z}_k)\big).$$

The gradient w.r.t. $\boldsymbol{x}_i$ is:

$$-\frac{\partial E}{\partial \boldsymbol{x}_i} = \sum_{k=1}^{K} \alpha_{i,k} \left(\boldsymbol{x}_i - f(\boldsymbol{z}_k)\right).$$

The gradient w.r.t. $\boldsymbol{z}_k$ is:

$$-\frac{\partial E}{\partial \boldsymbol{z}_k} = \sum_{i=1}^{N} \alpha_{i,k} \frac{\partial f(\boldsymbol{z}_k)}{\partial \boldsymbol{z}_k} \left(\boldsymbol{x}_i - f(\boldsymbol{z}_k)\right).$$

Note these are these are predictive coding updates, weighted by the attention $\alpha_{i,k}$.

Kronecker $\log \sum \exp$

We define a single set of child nodes $\{\boldsymbol{x}_i\}_{i=1}^{N}$ and *two* sets of parents:

$$\{\boldsymbol{z}_k\}_{k=1}^{K}, \quad \{\boldsymbol{u}_\ell\}_{\ell=1}^{L}.$$

We combine these factors *multiplicatively* via a *Kronecker* structure, giving a single $\log \sum \exp$ term over all pairs (k, ℓ).

Similarity: For each child $\boldsymbol{x}_i$ and each parent pair (k, ℓ), define

$$\mathrm{sim}\big(\boldsymbol{x}_i; \boldsymbol{z}_k, \boldsymbol{u}_\ell\big) = \mathrm{sim}_z(\boldsymbol{x}_i, \boldsymbol{z}_k) \; + \; \mathrm{sim}_u(\boldsymbol{x}_i, \boldsymbol{u}_\ell),$$

$$E^{\mathrm{Kron}} = -\sum_{i=1}^{N} \ln\Big(\sum_{k=1}^{K}\sum_{\ell=1}^{L} \exp\big(\mathrm{sim}(\boldsymbol{x}_i; \boldsymbol{z}_k, \boldsymbol{u}_\ell)\big)\Big).$$

Kronecker $\log \sum \exp$

The resulting attention is now a *joint* softmax over all parent pairs:

$$\alpha_{i,k,\ell} = \mathrm{softmax}_{k,\ell}\big(\mathrm{sim}(\boldsymbol{x}_i; \boldsymbol{z}_k, \boldsymbol{u}_\ell)\big)$$

$$= \frac{\exp\big(\mathrm{sim}(\boldsymbol{x}_i; \boldsymbol{z}_k, \boldsymbol{u}_\ell)\big)}{\sum_{k'=1}^{K} \sum_{\ell'=1}^{L} \exp\big(\mathrm{sim}(\boldsymbol{x}_i; \boldsymbol{z}_{k'}, \boldsymbol{u}_{\ell'})\big)}.$$

Gradients: For a child $\boldsymbol{x}_i$,

$$-\frac{\partial E}{\partial \boldsymbol{x}_i} = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \alpha_{i,k,\ell} \, \frac{\partial}{\partial \boldsymbol{x}_i} \mathrm{sim}(\boldsymbol{x}_i; \boldsymbol{z}_k, \boldsymbol{u}_\ell).$$

Similarly, for a parent $\boldsymbol{z}_k$:

$$-\frac{\partial E}{\partial \boldsymbol{z}_k} = \sum_{i=1}^{N} \sum_{\ell=1}^{L} \alpha_{i,k,\ell} \, \frac{\partial}{\partial \boldsymbol{z}_k} \mathrm{sim}_z(\boldsymbol{x}_i, \boldsymbol{z}_k),$$

Kronecker $\log \sum \exp$ (Discrete Case)

Setup: We define a single discrete child variable $\boldsymbol{x} \in \{1, \dots, D\}$ and two discrete parent variables:

$$\boldsymbol{z} \in \{1, \dots, K\}, \quad \boldsymbol{u} \in \{1, \dots, L\}.$$

The parameters $\boldsymbol{\theta} \in \mathbb{R}^{D \times K \times L}$ represent a joint categorical distribution.

Energy We define a single similarity function that depends on all three variables:

$$\mathrm{sim}(x, z, u) = \ln \theta_{x,z,u}.$$

$$E^{\mathrm{Kron}} = -\sum_{i=1}^{N} \ln\Big(\sum_{k=1}^{K} \sum_{\ell=1}^{L} \exp\big(\mathrm{sim}(x_i, z_k, u_\ell)\big)\Big).$$

Kronecker $\log \sum \exp$ (Discrete Case)

Tensor Attention: This induces a joint softmax over both parent variables:

$$\alpha_{i,k,\ell} = \frac{\exp\big(\mathrm{sim}(x_i, z_k, u_\ell)\big)}{\sum_{k'=1}^{K} \sum_{\ell'=1}^{L} \exp\big(\mathrm{sim}(x_i, z_{k'}, u_{\ell'})\big)}.$$

Interpretation:

- Parent variables (z_k, u_ℓ) conspire to jointly explain child variables x_i.
- The attention $\alpha_{i,k,\ell}$ generalizes categorical mixture models to factorial structures.

Multi $\log \sum \exp$

We define a set of child nodes $\{\boldsymbol{x}_i\}_{i=1}^N$ and a set of parent nodes $\{\boldsymbol{z}_k\}_{k=1}^K$. Each parent k is associated with multiple similarity terms, indexed by different factors.

Similarity functions: Each similarity term contributes independently to the energy function:

$$\mathrm{sim}_1(\boldsymbol{x}_i, \boldsymbol{z}_k), \quad \mathrm{sim}_2(\boldsymbol{x}_i, \boldsymbol{z}_k), \quad \ldots, \quad \mathrm{sim}_M(\boldsymbol{x}_i, \boldsymbol{z}_k).$$

Energy function:

$$E^{\mathrm{Multi}} = -\sum_{i=1}^{N} \ln\Bigg(\sum_{k=1}^{K} \exp\big(\textstyle\sum_{m=1}^{M} \mathrm{sim}_m(\boldsymbol{x}_i, \boldsymbol{z}_k)\big)\Bigg).$$

Here, all similarity terms are summed before the softmax, leading to a *joint* mixture model over multiple factors.

Multi $\log \sum \exp$

The resulting attention weight is:

$$\alpha_{i,k} = \mathrm{softmax}_k\left(\textstyle\sum_{m=1}^{M} \mathrm{sim}_m(\boldsymbol{x}_i, \boldsymbol{z}_k)\right).$$

With gradients:

$$-\frac{\partial E}{\partial \boldsymbol{x}_i} = \sum_{k=1}^{K} \alpha_{i,k} \sum_{m=1}^{M} \frac{\partial}{\partial \boldsymbol{x}_i} \mathrm{sim}_m(\boldsymbol{x}_i, \boldsymbol{z}_k).$$

$$-\frac{\partial E}{\partial \boldsymbol{z}_k} = \sum_{i=1}^{N} \alpha_{i,k} \sum_{m=1}^{M} \frac{\partial}{\partial \boldsymbol{z}_k} \mathrm{sim}_m(\boldsymbol{x}_i, \boldsymbol{z}_k).$$

This formulation is useful when multiple similarties contribute to the similarity between a child and parent.

Block-slot Attention

Setup: We define a set of child nodes $\{\boldsymbol{x}_i\}_{i=1}^{N}$, a set of slot parent nodes $\{\boldsymbol{z}_k\}_{k=1}^{K}$, and a set of memory parent nodes $\{\boldsymbol{m}_\mu\}_{\mu=1}^{M}$. The parameters $\boldsymbol{\theta} = \{\boldsymbol{W}_K, \boldsymbol{W}_Q\}$ consist of projection matrices.

Energy: Each child $\boldsymbol{x}_i$ is compared to slot parents $\boldsymbol{z}_k$ and memory parents $\boldsymbol{m}_\mu$:

$$\mathrm{sim}_z(\boldsymbol{x}_i, \boldsymbol{z}_k) = (\boldsymbol{W}_K \boldsymbol{x}_i)^\top (\boldsymbol{W}_Q \boldsymbol{z}_k),$$

$$\mathrm{sim}_m(\boldsymbol{x}_i, \boldsymbol{m}_\mu) = \boldsymbol{x}_i^\top \boldsymbol{m}_\mu.$$

$$E^{\mathrm{BlockSlot}} = -\sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \exp\big(\mathrm{sim}_z(\boldsymbol{x}_i, \boldsymbol{z}_k)\big) \right)$$

$$-\sum_{i=1}^{N} \ln \left(\sum_{\mu=1}^{M} \exp\big(\mathrm{sim}_m(\boldsymbol{x}_i, \boldsymbol{m}_\mu)\big) \right).$$

Energy Transformer

Setup: We define a set of nodes $\{\boldsymbol{x}_i\}_{i=1}^{N}$ and a set of memory nodes $\{\boldsymbol{m}_\mu\}_{\mu=1}^{M}$. The parameters $\boldsymbol{\theta} = \{\boldsymbol{W}^Q, \boldsymbol{W}^K\}$ consist of projection matrices for queries and keys.

Energy: Each node $\boldsymbol{x}_i$ attends to itself (self-attention) and to memory nodes $\boldsymbol{m}_\mu$ (Hopfield attention):

$$\mathrm{sim}_{\mathsf{self}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{W}^Q \boldsymbol{x}_i)^\top (\boldsymbol{W}^K \boldsymbol{x}_j),$$

$$\mathrm{sim}_{\mathsf{memory}}(\boldsymbol{x}_i, \boldsymbol{m}_\mu) = \boldsymbol{x}_i^\top \boldsymbol{m}_\mu.$$

$$E^{\mathrm{ET}} = -\sum_{i=1}^{N} \ln\left(\sum_{j=1}^{N} \exp\left(\mathrm{sim}_{\mathsf{self}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)\right) - \sum_{i=1}^{N} \ln\left(\sum_{\mu=1}^{M} \exp\left(\mathrm{sim}_{\mathsf{memory}}(\boldsymbol{x}_i, \boldsymbol{m}_\mu)\right)\right).$$

Atari Model

Setup: We define a sequence of child nodes $\{\boldsymbol{x}_t\}_{t=1}^T$, a sequence of latent parent nodes $\{\boldsymbol{z}_t\}_{t=1}^T$, and a set of mode parameters $\{\boldsymbol{A}_k, \boldsymbol{b}_k, \Sigma_k\}_{k=1}^K$. The parameters $\boldsymbol{\theta} = \{\boldsymbol{A}_k, \boldsymbol{b}_k, \Sigma_k\}$ define the mappings.

Similarity Functions: The child $\boldsymbol{x}_t$ is assigned to a mixture component $\boldsymbol{z}_t$ using:

$$\mathrm{sim}_{\mathsf{GMM}}(\boldsymbol{x}_t, \boldsymbol{z}_t) = -\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{z}_t)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_t - \boldsymbol{z}_t).$$

The latent child $\boldsymbol{z}_t$ is explained by a linear mixture model:

$$\mathrm{sim}_{\mathsf{LM}}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) = -\frac{1}{2}(\boldsymbol{z}_t - \boldsymbol{A}_k \boldsymbol{z}_{t-1} - \boldsymbol{b}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{z}_t - \boldsymbol{A}_k \boldsymbol{z}_{t-1} - \boldsymbol{b}_k).$$

Atari Model

Energy:

$$E^{\mathrm{Atari}} = -\sum_{t=1}^{T} \ln\left(\sum_{k=1}^{K} \exp\bigl(\mathrm{sim}_{\mathsf{GMM}}(\boldsymbol{x}_t, \boldsymbol{z}_t)\bigr)\right)$$

$$-\sum_{t=1}^{T-1} \ln\left(\sum_{k=1}^{K} \exp\bigl(\mathrm{sim}_{\mathsf{LM}}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})\bigr)\right).$$

The gradient for the latent parent $\boldsymbol{z}_t$ combines both responsibilities:

$$-\frac{\partial E}{\partial \boldsymbol{z}_t} = \sum_{k=1}^{K} \alpha_{t,k}\, \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_t - \boldsymbol{z}_t) + \sum_{k=1}^{K} \beta_{t,k}\, \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{z}_t - \boldsymbol{A}_k \boldsymbol{z}_{t-1} - \boldsymbol{b}_k).$$

Spatiotemporal Attention

Hierarchy: We define a hierarchy of L layers. Each layer l contains K_l latent variables, where each variable evolves over T time steps. Concretely, let:

$$\boldsymbol{x}_{k,t}^{(l)} \in \mathbb{R}^{D_l}$$

denote the k-th variable (or "slot") in layer l at time t, with D_l-dimensional representations. The number of variables K_l may differ across layers.

Structure:

- **Inter-layer (vertical):** Each variable $\boldsymbol{x}_{k,t}^{(l)}$ is influenced by variables from the layer below $(l-1)$ at the same time t.
- **Intra-layer (concurrent):** Variables within the same layer l interact at each time step t.
- **Temporal:** Each variable $\boldsymbol{x}_{k,t}^{(l)}$ is influenced by its own past states $\boldsymbol{x}_{k,t'}^{(l)}$ for $t' < t$.

Spatiotemporal Attention

Inter-layer (vertical):

$$\mathrm{sim}_v^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \big(\boldsymbol{W}_v^{K,(l)}\,\boldsymbol{x}\big)^\top \big(\boldsymbol{W}_v^{Q,(l)}\,\boldsymbol{x}'\big).$$

Intra-layer (concurrent):

$$\mathrm{sim}_c^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \big(\boldsymbol{W}_c^{Q,(l)}\,\boldsymbol{x}\big)^\top \big(\boldsymbol{W}_c^{K,(l)}\,\boldsymbol{x}'\big).$$

Temporal (past states):

$$\mathrm{sim}_t^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \big(\boldsymbol{W}_t^{Q,(l)}\,\boldsymbol{x}\big)^\top \big(\boldsymbol{W}_t^{K,(l)}\,\boldsymbol{x}'\big).$$

Spatiotemporal Attention

Inter-layer Energy (Slot attentiom):

$$E_v^{(l)} = -\sum_{t=1}^{T}\sum_{c=1}^{K_{l-1}} \ln\left(\sum_{k=1}^{K_l} \exp\left(\mathrm{sim}_v^{(l)}(\boldsymbol{x}_{c,t}^{(l-1)}, \boldsymbol{x}_{k,t}^{(l)})\right)\right).$$

Intra-layer Energy (Self attention):

$$E_c^{(l)} = -\sum_{t=1}^{T}\sum_{k=1}^{K_l} \ln\left(\sum_{\substack{k'=1\\k'\neq k}}^{K_l} \exp\left(\mathrm{sim}_c^{(l)}(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{k',t}^{(l)})\right)\right).$$

Temporal Energy (Casual self attention):

$$E_t^{(l)} = -\sum_{k=1}^{K_l}\sum_{t=2}^{T} \ln\left(\sum_{t'<t} \exp\left(\mathrm{sim}_t^{(l)}(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{k,t'}^{(l)})\right)\right).$$

Spatiotemporal Attention

$$-\frac{\partial E}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \underbrace{-\frac{\partial E_v^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}}}_{\text{bottom-up}} + \underbrace{-\frac{\partial E_v^{(l+1)}}{\partial \boldsymbol{x}_{k,t}^{(l)}}}_{\text{top-down}} + \underbrace{-\frac{\partial E_c^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}}}_{\text{intra-layer}} + \underbrace{-\frac{\partial E_t^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}}}_{\text{temporal}}.$$

Bottom-up gradient

$$-\frac{\partial E_v^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \sum_{c=1}^{K_{l-1}} \mathrm{softmax}_k\left(\boldsymbol{A}_{c,k}\right) \boldsymbol{W}_v^{Q,(l)\top} \boldsymbol{W}_v^{K,(l)} \boldsymbol{x}_{c,t}^{(l-1)}.$$

Top-down gradient

$$-\frac{\partial E_v^{(l+1)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \sum_{p=1}^{K_{l+1}} \mathrm{softmax}_p\left(\boldsymbol{A}_{k,p}\right) \boldsymbol{W}_v^{K,(l+1)\top} \boldsymbol{W}_v^{Q,(l+1)} \boldsymbol{x}_{p,t}^{(l+1)}.$$

Spatiotemporal Attention

Intra-layer gradient

$$-\frac{\partial E_c^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \sum_{k' \neq k} \mathrm{softmax}_k \left(\boldsymbol{A}_{k',k} \right) \boldsymbol{W}_c^{K,(l)\top} \boldsymbol{W}_c^{Q,(l)} \boldsymbol{x}_{k',t}^{(l)}$$

$$+ \sum_{k' \neq k} \mathrm{softmax}_{k'} \left(\boldsymbol{A}_{k,k'} \right) \boldsymbol{W}_c^{Q,(l)\top} \boldsymbol{W}_c^{K,(l)} \boldsymbol{x}_{k',t}^{(l)}.$$

Temporal gradient

$$-\frac{\partial E_t^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \sum_{t' < t} \mathrm{softmax}_{t'} \left(\boldsymbol{A}_{t,t'} \right) \boldsymbol{W}_t^{Q,(l)\top} \boldsymbol{W}_t^{K,(l)} \boldsymbol{x}_{k,t'}^{(l)}.$$

Remark: By merging inter-layer, intra-layer, and temporal connections into a single large log-sum-exp term, we increase flexibility but at the cost of $\mathcal{O}((N+K)\,T)^2$ complexity.

Renormalised mixture models

We define a hierarchy of L layers, each containing K_l latent variables of dimension D_l, where K_l decreases with l $(K_1 > K_2 > \cdots > K_L)$.

Each variable $\boldsymbol{x}_{k,t}^{(l)}$ at layer l receives input only from a local subset of variables in layer $l-1$, forming a *receptive field*:

$$\mathcal{R}_k^{(l)} \subseteq \{1, \ldots, K_{l-1}\}.$$

In this setting, intra-layer and temporal energy remain the same.

Renormalised Mixture Model

Each variable $\boldsymbol{x}_{k,t}^{(l)}$ at layer l is explained only by a local receptive field $\mathcal{R}_k^{(l)}$ in layer $l-1$:

$$E_v^{(l)} = -\sum_{t=1}^{T}\sum_{k=1}^{K_l} \ln\left(\sum_{c\in\mathcal{R}_k^{(l)}} \exp\left(\mathrm{sim}_v^{(l)}(\boldsymbol{x}_{c,t}^{(l-1)}, \boldsymbol{x}_{k,t}^{(l)})\right)\right).$$

Define:

$$\boldsymbol{A}_{c,k} = \mathrm{sim}_v^{(l)}(\boldsymbol{x}_{c,t}^{(l-1)}, \boldsymbol{x}_{k,t}^{(l)}), \quad \boldsymbol{A}_{k,p} = \mathrm{sim}_v^{(l+1)}(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{p,t}^{(l+1)}).$$

Then the bottom-up and top-down gradients are:

$$-\frac{\partial E_v^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \sum_{c\in\mathcal{R}_k^{(l)}} \mathrm{softmax}_c\left(\boldsymbol{A}_{c,k}\right)\frac{\partial}{\partial \boldsymbol{x}_{k,t}^{(l)}}\mathrm{sim}_v^{(l)}(\boldsymbol{x}_{c,t}^{(l-1)}, \boldsymbol{x}_{k,t}^{(l)}),$$

$$-\frac{\partial E_v^{(l+1)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \sum_{p\in\mathcal{R}_k^{(l+1)}} \mathrm{softmax}_p\left(\boldsymbol{A}_{k,p}\right)\frac{\partial}{\partial \boldsymbol{x}_{k,t}^{(l)}}\mathrm{sim}_v^{(l+1)}(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{p,t}^{(l+1)}).$$

Bayesian model expansion

If a new data point $\boldsymbol{x}_{N+1}$ is not well explained by existing parent components, we introduce a new parent to explain it.
We evaluate the energy contribution of the new data point:

$$E_{N+1} = -\ln\left(\sum_{k=1}^{K} \exp\big(\mathrm{sim}(\boldsymbol{x}_{N+1}, \boldsymbol{z}_k,)\big)\right).$$

If $E_{N+1} > \tau$ (for some threshold τ), then $\boldsymbol{x}_{N+1}$ is not sufficiently explained, and we add a new parent component $\boldsymbol{z}_{K+1}$, where the new parent is initialized based on $\boldsymbol{x}_{N+1}$.

Coordinate Ascent

Energy Function:
We consider an energy of the form

$$E(\{\boldsymbol{x}_j\}, \{\boldsymbol{\mu}_i\}; \theta) = -\sum_{j=1}^{N} \ln\Big(\sum_{i=1}^{S} \exp\big(\mathrm{sim}_\theta(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\big)\Big),$$

where θ denotes the parameters of the similarity function.

EM Procedure:

- **E-step:** Update latent variables $\{\boldsymbol{\mu}_i\}$ given fixed θ.
- **M-step:** Update parameters θ in closed form given current slots.

Coordinate Ascent

Slot Attention Example:

$$\theta = \{\boldsymbol{W}_K, \boldsymbol{W}_Q\}, \quad \mathrm{sim}_\theta(\boldsymbol{x}_j, \boldsymbol{\mu}_i) = (\boldsymbol{W}_K \boldsymbol{x}_j)^\top (\boldsymbol{W}_Q \boldsymbol{\mu}_i).$$

Define the attention matrix $\mathbf{A}$ with entries

$$A_{ji} = \mathrm{softmax}_i\big((\boldsymbol{W}_K \boldsymbol{x}_j)^\top (\boldsymbol{W}_Q \boldsymbol{\mu}_i)\big).$$

Goal: Perform an **E-step** (update slots $\boldsymbol{\mu}_i$) and **M-step** (update $\boldsymbol{W}_K, \boldsymbol{W}_Q$).

Coordinate Ascent

M-step: Update Parameters
Given fixed slots $\{\boldsymbol{\mu}_i\}$ and attention $\mathbf{A}$, we update $\boldsymbol{W}_K, \boldsymbol{W}_Q$. We consider minimizing

$$\min_{\boldsymbol{W}_K, \boldsymbol{W}_Q} \sum_{j=1}^{N} \sum_{i=1}^{S} A_{ji} \left\| \boldsymbol{W}_K \boldsymbol{x}_j - \boldsymbol{W}_Q \boldsymbol{\mu}_i \right\|^2.$$

Setting the gradient to zero yields a weighted least-squares problem in $\boldsymbol{W}_K$ and $\boldsymbol{W}_Q$.

Key idea:

- $\mathbf{A}$ is fixed (like responsibilities in EM).
- Solve for $\boldsymbol{W}_K, \boldsymbol{W}_Q$ in closed form.

Coordinate Ascent

Closed-Form Updates
Differentiate and set to zero:

$$\boldsymbol{W}_K = \Big(\sum_{j,i} A_{ji}\, \boldsymbol{W}_Q \boldsymbol{\mu}_i\, \boldsymbol{x}_j^\top\Big)\Big(\sum_{j,i} A_{ji}\, \boldsymbol{x}_j \boldsymbol{x}_j^\top\Big)^{-1}.$$

$$\boldsymbol{W}_Q = \Big(\sum_{j,i} A_{ji}\, \boldsymbol{W}_K \boldsymbol{x}_j\, \boldsymbol{\mu}_i^\top\Big)\Big(\sum_{j,i} A_{ji}\, \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top\Big)^{-1}.$$

Algorithm: Alternate E-step (update $\{\boldsymbol{\mu}_i\}$) and M-step (update $\boldsymbol{W}_K, \boldsymbol{W}_Q$) until convergence.

Open questions

- Hierarchical POMDPs and hierarchical mixture models differ only in their similairty functions.
- Crucially, hierarchical mixture models induce a discrete state space (in terms of their attention matrices).
- This is seen in SLDS, where we model the dynamics of the induced discrete state space (which mode is active at a given time) as a HMM.
- Is there a way to combine the two models, with continuous mixture models on the bottom and discrete mixture models on the top?