Attention via $\log \sum \exp$ Energy

Alexander Tschantz

January 21, 2025

1 General Framework

Setup. We consider a single set of nodes $\boldsymbol{v} = \{\boldsymbol{v}_a : a \in \{1, 2, \ldots, A\}\}$, where each node $\boldsymbol{v}_a \in \mathbb{R}^d$. The relationships between these nodes are defined by a set of M energy functions $\{E_m : m \in \{1, 2, \ldots, M\}\}$. Each energy function E_m defines a subset of nodes acting as *children* $C_m \subseteq \{1, 2, \ldots, A\}$ and a subset acting as *parents* $P_m \subseteq \{1, 2, \ldots, A\}$, which may overlap.

Energy. Each energy function E_m defines a similarity function:

$$\operatorname{sim}\big(\boldsymbol{v}_c, \boldsymbol{v}_p\big) \quad : \quad \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \tag{1}$$

which produces a scalar similarity between a child $\boldsymbol{v}_c$ and a parent $\boldsymbol{v}_p$. Using $\{\boldsymbol{v}_c\} = \{\boldsymbol{v}_c : c \in C_m\}$ and $\{\boldsymbol{v}_p\} = \{\boldsymbol{v}_p : p \in P_m\}$, the energy for E_m is defined as:

$$E_m\big(\{\boldsymbol{v}_c\}, \{\boldsymbol{v}_p\}\big) \;=\; -\sum_{c \in C_m} \ln\Big(\sum_{p \in P_m} \exp\big(\operatorname{sim}(\boldsymbol{v}_c, \boldsymbol{v}_p)\big)\Big). \tag{2}$$

The global energy sums over all energy functions:

$$E\big(\{\boldsymbol{v}\}\big) \;=\; \sum_{m=1}^{M} E_m\big(\{\boldsymbol{v}_c\}, \{\boldsymbol{v}_p\}\big). \tag{3}$$

Gradient Updates. For a single node $\boldsymbol{v}_a$, the gradient of the global energy E w.r.t. $\boldsymbol{v}_a$ decomposes into two terms. Let $\mathcal{M}_c(a) = \{m : a \in C_m\}$ denote the energy functions where $\boldsymbol{v}_a$ acts as a *child*, and $\mathcal{M}_p(a) = \{m : a \in P_m\}$ the energy functions where $\boldsymbol{v}_a$ acts as a *parent*. Then:

$$\begin{aligned} -\frac{\partial E}{\partial \boldsymbol{v}_a} \;=\; & \underbrace{\sum_{m \in \mathcal{M}_c(a)} \sum_{p \in P_m} \operatorname{softmax}_p\Big(\operatorname{sim}(\boldsymbol{v}_a, \boldsymbol{v}_p)\Big) \frac{\partial}{\partial \boldsymbol{v}_a} \operatorname{sim}\big(\boldsymbol{v}_a, \boldsymbol{v}_p\big)}_{\boldsymbol{v}_a \text{ acting as a child}} \\ & + \underbrace{\sum_{m \in \mathcal{M}_p(a)} \sum_{c \in C_m} \operatorname{softmax}_a\Big(\operatorname{sim}(\boldsymbol{v}_c, \boldsymbol{v}_a)\Big) \frac{\partial}{\partial \boldsymbol{v}_a} \operatorname{sim}\big(\boldsymbol{v}_c, \boldsymbol{v}_a\big)\,.}_{\boldsymbol{v}_a \text{ acting as a parent}} \end{aligned} \tag{4}$$

The first term captures contributions from $\boldsymbol{v}_a$ being explained by its parents, while the second term captures contributions from $\boldsymbol{v}_a$ explaining its children.

2 Gaussian Mixture Models (GMMs)

Setup. We have N data points (children) $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \in C = \{1, \ldots, N\}$, and K mixture components (parents), each with mean $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma}_k$, $k \in P = \{1, \ldots, K\}$. Let π_k be the mixing proportion.

Similarity function. We define

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

Energy.

$$E^{\text{GMM}}\Big(\{\boldsymbol{x}_i\},\{\boldsymbol{\mu}_k\}\Big) \;=\; -\sum_{i=1}^{N}\ln\Big(\sum_{k=1}^{K}\exp\big(\text{sim}(\boldsymbol{x}_i,\boldsymbol{\mu}_k)\big)\Big). \tag{5}$$

Gradients. If we differentiate w.r.t. $\boldsymbol{\mu}_k$, then

$$-\frac{\partial E^{\text{GMM}}}{\partial \boldsymbol{\mu}_k} \;=\; \sum_{i=1}^{N}\text{softmax}_k\big(\boldsymbol{A}_{ik}\big)\;\boldsymbol{\Sigma}_k^{-1}\big(\boldsymbol{x}_i-\boldsymbol{\mu}_k\big).$$

Setting this gradient to zero yields the usual GMM M-step:

$$\boldsymbol{\mu}_k \;=\; \frac{\sum_{i=1}^{N}\text{softmax}_k\big(\boldsymbol{A}_{ik}\big)\;\boldsymbol{x}_i}{\sum_{i=1}^{N}\text{softmax}_k\big(\boldsymbol{A}_{ik}\big)}.$$

3 Cross Attention

Setup. We have a set of child vectors (queries) $\boldsymbol{Q}\in\mathbb{R}^{d\times N_Q}$ and a set of parent vectors (keys) $\boldsymbol{K}\in\mathbb{R}^{d\times N_K}$. Let

$$C \;=\; \{1,\ldots,N_Q\},\quad P \;=\; \{1,\ldots,N_K\},$$

so $\boldsymbol{v}_c=\boldsymbol{q}_c$ is the c-th query, and $\boldsymbol{v}_p=\boldsymbol{k}_p$ is the p-th key. Suppose we have learnable weight matrices $\boldsymbol{W}^Q,\boldsymbol{W}^K\in\mathbb{R}^{d\times d}$. Then

$$\boldsymbol{q}_c \;=\; \boldsymbol{W}^Q\boldsymbol{x}_c^Q,\quad \boldsymbol{k}_p \;=\; \boldsymbol{W}^K\boldsymbol{x}_p^K,$$

where $\boldsymbol{x}_c^Q$ is the raw c-th query token and $\boldsymbol{x}_p^K$ the raw p-th key token.

Similarity function.

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

Energy.

$$E^{\text{Cross}}\Big(\{\boldsymbol{q}_c\},\{\boldsymbol{k}_p\}\Big) \;=\; -\sum_{c=1}^{N_Q}\ln\Big(\sum_{p=1}^{N_K}\exp\big(\boldsymbol{q}_c^\top\,\boldsymbol{k}_p\big)\Big). \tag{6}$$

Gradients.

$$-\frac{\partial E^{\text{Cross}}}{\partial \boldsymbol{q}_c} \;=\; \sum_{p=1}^{N_K}\text{softmax}_p\Big(\boldsymbol{q}_c^\top\boldsymbol{k}_p\Big)\,\boldsymbol{k}_p. \tag{7}$$

$$-\frac{\partial E^{\text{Cross}}}{\partial \boldsymbol{k}_p} \;=\; \sum_{c=1}^{N_Q}\text{softmax}_p\Big(\boldsymbol{q}_c^\top\boldsymbol{k}_p\Big)\,\boldsymbol{q}_c. \tag{8}$$

When mapping back to the raw tokens $\boldsymbol{x}_c^Q$ or $\boldsymbol{x}_p^K$, chain-rule multiplies by $\boldsymbol{W}^Q$ or $\boldsymbol{W}^K$, respectively.

4 Hopfield Networks

Setup. We have a set of *children* data vectors $\boldsymbol{x}_i\in\mathbb{R}^d$, $i\in C=\{1,\ldots,N\}$, and a set of *parent* memory vectors $\boldsymbol{m}_\mu\in\mathbb{R}^d$, $\mu\in P=\{1,\ldots,K\}$.

Similarity function.

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

Energy.

$$E^{\text{Hopfield}}\Big(\{\boldsymbol{x}_i\},\{\boldsymbol{m}_\mu\}\Big) \;=\; -\sum_{i=1}^{N}\ln\Big(\sum_{\mu=1}^{K}\exp\big(\boldsymbol{x}_i^\top\boldsymbol{m}_\mu\big)\Big). \tag{9}$$

Gradients.

$$-\frac{\partial E^{\text{Hopfield}}}{\partial \boldsymbol{x}_i} \;=\; \sum_{\mu=1}^{K} \operatorname{softmax}_\mu\big(\boldsymbol{x}_i^\top \boldsymbol{m}_\mu\big)\, \boldsymbol{m}_\mu. \tag{10}$$

$$-\frac{\partial E^{\text{Hopfield}}}{\partial \boldsymbol{m}_\mu} \;=\; \sum_{i=1}^{N} \operatorname{softmax}_\mu\big(\boldsymbol{x}_i^\top \boldsymbol{m}_\mu\big)\, \boldsymbol{x}_i. \tag{11}$$

5 Slot Attention

Setup. Let $\boldsymbol{x}_j \in \mathbb{R}^d$, $j \in C = \{1, \ldots, N\}$ be the children (tokens), and $\boldsymbol{\mu}_i \in \mathbb{R}^d$, $i \in P = \{1, \ldots, S\}$ be the parents (slots). We typically apply linear transforms $\boldsymbol{W}_K, \boldsymbol{W}_Q \in \mathbb{R}^{d\times d}$ to form

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

Energy.

$$E^{\text{Slot}}\Big(\{\boldsymbol{x}_j\}, \{\boldsymbol{\mu}_i\}\Big) \;=\; -\sum_{j=1}^{N} \ln\Big(\sum_{i=1}^{S} \exp\big(\operatorname{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\big)\Big). \tag{12}$$

Gradients.

$$-\frac{\partial E^{\text{Slot}}}{\partial \boldsymbol{x}_j} \;=\; \sum_{i=1}^{S} \operatorname{softmax}_i\Big(\operatorname{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\Big)\, \boldsymbol{W}_K^\top \boldsymbol{W}_Q\, \boldsymbol{\mu}_i. \tag{13}$$

$$-\frac{\partial E^{\text{Slot}}}{\partial \boldsymbol{\mu}_i} \;=\; \sum_{j=1}^{N} \operatorname{softmax}_i\Big(\operatorname{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\Big)\, \boldsymbol{W}_Q^\top \boldsymbol{W}_K\, \boldsymbol{x}_j. \tag{14}$$

6 Self-Attention

Setup. In self-attention, every node can act as both a child (query) and a parent (key). Concretely, let us have N tokens $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. We form

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

for $i = 1, \ldots, N$. Thus the set $C = \{1, \ldots, N\}$ and $P = \{1, \ldots, N\}$ coincide, with

$$\operatorname{sim}\big(\boldsymbol{x}_c, \boldsymbol{x}_p\big) \;=\; \big(\boldsymbol{W}^Q \boldsymbol{x}_c\big)^\top \big(\boldsymbol{W}^K \boldsymbol{x}_p\big).$$

Energy.

$$E^{\text{SA}}\Big(\{\boldsymbol{x}_i\}\Big) \;=\; -\sum_{c=1}^{N} \ln\Big(\sum_{p=1}^{N} \exp\Big((\boldsymbol{W}^Q \boldsymbol{x}_c)^\top (\boldsymbol{W}^K \boldsymbol{x}_p)\Big)\Big). \tag{15}$$

Gradients. Since each $\boldsymbol{x}_i$ is *both* a child and a parent, its gradient is a sum of two terms (the child side and the parent side). Writing it out explicitly:

$$\begin{aligned} -\frac{\partial E^{\text{SA}}}{\partial \boldsymbol{x}_i} \;=\; & \underbrace{\sum_{p=1}^{N} \operatorname{softmax}_p\Big((\boldsymbol{W}^Q \boldsymbol{x}_i)^\top (\boldsymbol{W}^K \boldsymbol{x}_p)\Big)\, \boldsymbol{W}_Q^\top \boldsymbol{W}_K\, \boldsymbol{x}_p}_{\text{child } i \text{ being explained by parents } p} \\ & + \underbrace{\sum_{c=1}^{N} \operatorname{softmax}_i\Big((\boldsymbol{W}^Q \boldsymbol{x}_c)^\top (\boldsymbol{W}^K \boldsymbol{x}_i)\Big)\, \boldsymbol{W}_K^\top \boldsymbol{W}_Q\, \boldsymbol{x}_c}_{\text{parent } i \text{ explaining children } c}. \end{aligned} \tag{16}$$